Investigating human population structure through time with new computational methods and ancient DNA data

Dissertation

zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)

Vorgelegt dem Rat der
Biologisch-Pharmazeutischen Fakultät der
Friedrich-Schiller-Universität Jena

Von Dipl. Genetik Ke Wang
Geb. Am 22.11.1996 in Heze

Gutachter:

1. Dr. Stephan Schiffels (Max-Planck-Institue für Menschheitsgeschichte, Jena)

2. Prof. Dr. Matthias Steinrücken (University of Chicago, Chicago)

3. Prof. Dr. Christina Warinner (Friedrich-Schiller-Universität, Jena)

Dissertation eingereicht am: 20. April 2020
Tag der öffentlichen Verteidigung: 14. Januar 2021

Table of Contents

1. Introduction

Where do we come from? What are we? Where are we going? It is difficult to predict where we are going, but our genomes can tell where we come from and what we are. The rise of next-generation sequencing (NGS) technologies makes sequencing the whole human genome increasingly affordable. The emergence of efficient genome assembly tools (Li et al., 2009; Li & Durbin, 2011) simplifies the post-sequencing process tremendously. The first application of NGS to the whole human genome identified genetic variations across the entire sequence, such as single nucleotide polymorphisms (SNPs) (Wheeler et al., 2008). Genetic variation is defined as the difference in DNA across individual sequences. Such difference across species/populations/individuals explains "what we are" genetically. Meanwhile these variations are the key to study "where do we come from", i.e. the past human demographic history (Haak et al., 2015; Lazaridis et al., 2014, 2016; Mathieson et al., 2015). Several large-scale modern human genome projects, such as the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015) and the Simons Genome Diversity Project (SGDP) (Mallick et al., 2016), provide us with a comprehensive description on present-day human genome variation, and also enables us to investigate the global picture of human separation, movement and admixture among worldwide populations. In manuscript A of this thesis, I present a new inference framework called MSMC-IM on estimating the key human demographic feature - population separation, via quantified gene flow across populations. By applying this new tool to the SGDP dataset, I trace the process of human genetic diversification on a global scale, and in particular track the genetic footprints of admixture from a deeply diverged unknown population in present-day African populations like San and Mbuti.

To answer "where do we come from", an alternative way is to study the genomic data of our ancestors (ancient individuals) directly, instead of extrapolating information from modern human genomes. The first ancient DNA (aDNA) studies from 1980s, used bacterial cloning technique and later the targeted DNA amplification - Polymerase Chain Reaction (PCR), to retrieve aDNA from an Egyptian mummy (S. Pääbo, 1985) and quagga (Higuchi et al., 1984). Compared to modern DNA, studying aDNA has two major obstacles to overcome: i) fragmented DNA sequences (S. Pääbo, 1989), ii) contamination from the preservation environment (Zischler et al., 1995). The threat of high environmental DNA in ancient specimens, adds the difficulty of extracting authentic human DNA and the cost of sequencing meanwhile. The former limits the power of PCR which only targets relatively long DNA fragments while the average aDNA fragments are shorter than 100 bp (Sawyer et al., 2012), and complicates the authentication of aDNA by mistakenly amplifying long modern contaminating DNA fragments (Krause, Fu, et al., 2010). The contamination introduced from environment and PCR amplification remains an issue, although strict authentication criteria have since been developed over decades (Cooper & Poinar, 2000; Svante Pääbo et al., 2004).

The high sequencing throughput of NGS brings the first reformation to the archaeogenetics field. It allows parallel sequencing of DNA molecules (Metzker, 2010), and more importantly, allows the 'reading' of the complete sequence of DNA fragments with short length, making it possible to authenticate aDNA through the pattern of postmortem damage - nucleotide deamination, and fragmentation patterns (Krause, Briggs, et al., 2010). The second technological revolution for archaeogenetics is DNA enrichment techniques. To study

genetic variations in the genome, In-solution DNA enrichment ('capture') can be used for boosting the sequencing efficiency on single nucleotide polymorphisms (SNPs) via specifically designed probes targeting these SNPs of interest (Briggs et al., 2009; Burbano et al., 2010; Maricic et al., 2010). The list of SNPs designed for capture are selected informative sites from either mitochondria (Briggs et al., 2009; Fu et al., 2013; Maricic et al., 2010) or autosomes (Haak et al., 2015; Mathieson et al., 2015). The total number of autosomal SNPs in capture has expanded from 390k (Haak et al., 2015) to 1240k SNPs (Mathieson et al., 2015) ('1240k capture'), with the latter now widely used in aDNA studies (Lazaridis et al., 2016; Mittnik et al., 2019; Posth et al., 2018; Skoglund et al., 2017). Meanwhile, sampling techniques have also improved. In particular, petrous bones (Pinhasi et al., 2015) and teeth (Hansen et al., 2017) have been identified as anatomical elements with good preservation of aDNA. Together with the 1240k capture technique, these technical advancements allow us to recover aDNA from very old skeletons of low endogenous DNA content (i.e. proportion of authentic human DNA out of all amplified DNA sequences) and severe DNA degradation. In manuscript B and C of this thesis, I present two aDNA studies on i) 20 ancient individuals from sub-Saharan Africa dated to between 4000 BP to 300 BP, ii) 214 ancient individuals from the Eastern Steppe zone i.e. Mongolia and the neighbouring northern region up to Lake Baikal spanning from ca. 4600BC until 1400AD.

1.1 Learning population demographic history from genomic data

Present-day population structure is shaped by a series of complex demographic events, involving population splitting, gene flow exchange after divergence, and population size changes. In the field of population genetics, many methods aim to fit some type of demographic model to genetic data so as to make inferences on the key parameters of population history such as split time, effective population size, migration rate and admixture proportions. To study the split time and gene flow between populations, currently there are two main approaches - one based on the allele frequency spectrum (AFS) (Excoffier et al., 2013; Gutenkunst et al., 2009), and the other based on the sequentially Markov coalescent (SMC) (Marjoram & Wall, 2006; McVean & Cardin, 2005). AFS-based methods utilize the allele count distribution of millions of SNPs across populations to compute the likelihood under a pre-assumed divergence model (Sousa & Hey, 2013). The AFS approach has the advantage that it can be applied to complex demographic models with more than two populations (J. A. Kamm et al., 2017), although the computational costs become heavy when the sample size increases. But the main drawback is it assumes all SNPs are independent, which discards all local linkage equilibrium (LD) patterns, which in reality the data does contain. Consequently, AFS-based methods study demographic processes like gene flow or admixture, without using all available data and especially ignoring LD patterns (Sousa & Hey, 2013) despite the advantage of being applied to multiple populations. The SMC-based approach, in contrast, takes advantage of LD information among loci for demographic parameter inferences, which allows to extract extra information than AFS-based methods.

The SMC scheme, first proposed by McVean and Cardin (McVean & Cardin, 2005), employs phased DNA segments (haplotype) information and assumes free recombination and linkage among loci, served as a simplified model of the standard coalescent process. In a basic coalescent with recombination model, every recombination event generates new genealogies across sampled extant sequences, when moving spatially along the genome

(Wiuf & Hein, 1999). When the recombination rate is high, the possible ancestral recombination graphs numerically expands, resulting from increasing separations of the genealogical processes. The SMC reduces the complexity substantially by modelling the sequential generation of genealogies as a Markovian model along a chromosome (McVean & Cardin, 2005). Because a Markov process describes a sequence of possible events in continuous-time, where an event happens at a probability depending on the state attained in the previous event, it has similar properties to the distribution of a coalescent tree of genealogies along a sequence, which depends on the state from the previous recombination event (McVean & Cardin, 2005). The derived/specialized formats of SMC have been widely used to whole genome sequence data for various inferences so far. In this chapter I will provide an overview on SMC-based methods and inferred human history from their application to real human data.

1.1.1 Whole genome sequences help on reconstructing population history

The history of our common ancestors is embedded in the whole human genome sequence. To know about the past, reconstructing the distribution of the time to the most recent common ancestor (TMRCA) between two sequences is critical. A pair of human sequences must be either from the same individual since a human has a diploid genome with one set of chromosomes from the father and the other from the mother, or from two different individuals within or across populations. In the former case, we can learn about the size change in the population represented by a single individual (Li & Durbin, 2011), while in the latter we can learn about the differentiations across genetically different populations (Schiffels & Durbin, 2014a). For a pair of popuations or even more, alleles sampled at every single allelic site have their own TMRCA as a result of multiple evolutionary forces in the past. Recombination events seperate one TRMCA from another along the genomes we sampled. We can model these discrete TMRCA distributions along one or multiple whole genome sequences by employing some properties of SMC framework (Li & Durbin, 2011; Schiffels & Durbin, 2014a; Sheehan et al., 2013; Steinrücken et al., 2019), and make estimates on how fast the DNA segments have coalesced, and what demographic events in the past may have caused this.

1.1.2 Applications of the sequential Markov coalescent for demographic inference

Li and Durbin proposed the pairwise sequentially Markovian coalescent (PSMC) model (Li & Durbin, 2011) to tackle the question of population size changes based on a single diploid individual. As a special case of the SMC model, this method only focuses on two haplotypes from one diploid individual. It utilizes a Hidden Markov Model (HMM) - the sequential state in a Markov process that is only partially observable - to estimate the local TMRCA on a continuous-time level. The observation in this HMM is the density of heterozygous sites, and the hidden states are discretized TMRCAs. The transition between states are resulted from ancestral recombinations. The PSMC model has wide application for inferring population size changes from a single diploid individual of reliable diploid consensus calls, without requiring a phased genome (i.e. assigning alleles to peternal and maternal chromosomes). However, the estimates from PSMC are limited to between 20 kyr ago and 3 myr ago since beyond this time scope, only a few recombination events are left for two sequences. Extending the PSMC model to multiple sequences may therefore resolve the time scope limitation.

Later, Schiffels and Durbin proposed the multiple sequentially Markovian coalescent (MSMC) model (Schiffels & Durbin, 2014a), to address more complex demographic questions e.g. population separation. MSMC analyses multiple sequences simultaneously, focusing on the first coalescent events between any pair of sequences, to allow higher resolution in recent times. Because it is computationally heavy to enumerate all local genealogical trees under the PSMC framework for more than two haplotypes, MSMC implements a new algorithm by taking only the first coalescence between any two haplotypes and the total branch length in the genealogy tree as the hidden states. Therefore, in theory, the resolution of MSMC in recent times increases with increasing number of haplotypes sampled, as the time to the first coalescence event decreases when more haplotypes are sampled. While in reality, it is at the cost of high computational requirements. The most important novelty of MSMC is making inferences when two populations split based on the relative cross coalescence rate (rCCR), which is the cross-population coalescence rate divided by the average within-population coalescence rate. The time point when rCCR hits 0.5 is usually interpreted as the heuristic estimate of split time. Given that either within- or cross-population coalescence rates are time dependent in a continuous time scale, the rCCR also changes along time continuously, which actually suggests the possibility of interpreting continuous population separation process from the rCCR curves, rather than the pulse-based split.

MSMC2 is the successor of MSMC but with improvements on the algorithm and computational efficiency. It was first introduced in the Supplement of Ref (Malaspinas et al., 2016), and formally published in Ref (K. Wang et al., 2020). It uses pairwise HMMs for all pairs of sampled haplogroups, which takes the full distribution of pairwise TMRCAs into account, rather than only the first coalescent event across all pairs, and computes the overall likelihood by simply multiplying across all pairs as a composite likelihood. The way of calculating composite likelihood ignores the correlations of hidden states across pairs, which in practice avoids biases suffered from increasing number of haplotypes so as to have good resolution in ancient times in comparison to MSMC. Similar to MSMC, MSMC2 also requires phased haplotypes as input for reliable cross-population coalescence rate estimates, while when used for within-population coalescence rate estimate only, MSMC2 works the same way as PSMC without requirement on phasing. For both MSMC and MSMC2, population separations are interpreted only from the midpoint of rCCR curves, which is heuristic, although many key aspects of population separation, such as post-split migration and archaic introgression, are encrypted in these curves starting from 0 and ending at 1.

Since interpreting rCCR is always hypothesis-free, fitting an explicit model that is more parameterized would help us in understanding the population separation process in a simpler and more straightforward way. In manuscript A of this thesis, we propose a new approach MSMC-IM, to measure the separation and migration between a pair of populations quantitatively. MSMC-IM fits a continuous Isolation-Migration (IM) model to the distribution of coalescence times estimated from MSMC/MSMC2's piecewise constant model, and therefore maintains the continuous character of population separation from MSMC without explicitly specifying a complex population phylogeny. The continuous IM model is explicitly defined by a piecewise constant symmetric migration rate across populations, and piecewise constant population size within each population, discarding the concept of ancestral populations and split time in a classic two-population IM model (Y. Wang & Hey, 2010).

Hobolth and colleagues (Hobolth et al., 2011) proposed that the distribution of coalescent time density under a classic two-population IM model can be computed from a continuous Markov process, providing a theoretical basis for fitting the continuous IM model to the MSMC/MSMC2 model. Overall, this novel approach, MSMC-IM, allows us to quantitatively decode the complex population separation processes with the help of a structured simple IM model.

1.1.3 A brief overview on the human demographic history

Various SMC-based methods, e.g. PSMC (Li & Durbin, 2011), MSMC (Schiffels & Durbin, 2014a), diCal and diCal2 (Sheehan et al., 2013; Steinrücken et al., 2019), and SMC++ (Terhorst et al., 2017), have been applied to high-coverage whole genome human data to reconstruct two key features of human demography: i) the history of effective population size, and ii) the dynamics of population separation. The latter also includes small-scale gene flow after population split, as well as the main population split process. Overall, it is critical to understand how people migrate in the past, and how the post-split admixture shaped present-day population structure. In this section, I will give a brief summary on two features of human demography inferred from these SMC-based methods.

Looking backward in time, all populations show a similar history of population size changes before 300,000 years ago (Schiffels & Durbin, 2014a; Terhorst et al., 2017). From 300,000 years, African and non-African populations started going through a population decline process, though the severity of the bottleneck was different. African populations experienced mild reduction in effective population size with an extended time period followed by gradual growth from 100,000 to 10,000 years ago, while non-African populations showed a much steeper decline until about 50,000 years ago after which the population size rapidly increases until 10,000 years ago (Schiffels & Durbin, 2014a; Terhorst et al., 2017). The divergence time between African (using Yoruba as a representative) and non-African populations (using central European as a representative) is estimated to be around 60,000 to 80,000 years ago (Schiffels & Durbin, 2014a), roughly corresponding to the time point when the bottleneck of non-African populations is most severe. Among non-African populations, the separation between European and East Asian (using Han Chinese as a representative) is estimated to be around 20,000 and 40,000 years ago, followed by the separation between East Asian and American (using Mexican as a representative) at around 20,000 years ago. Expectedly, the youngest separation occurred within continents: 8,000-9,000 years ago between Han Chinese and Japanese, 5,000-6,000 years ago between central Europeans and southern Europeans (Schiffels & Durbin, 2014a). However the separations within African population are more likely to be more ancient, accompanied with gradual processes instead of a clean split (Schiffels & Durbin, 2014a).

Population separation is a complex process, as populations are often admixed with other populations after the main separation. There are methods like diCal2 (Steinrücken et al., 2019) and G-Phocs (Gronau et al., 2011) which address this question using a structured model with strong assumptions on the demographic events among existing modern-day populations. In addition, methods like S* (Plagnol & Wall, 2006) and Sprime (Browning et al., 2018) are specifically designed for detecting admixture from distinct archaic populations (e.g. Neandertal or Denisovan). Under these explicitly defined demographic models, diCal2

estimates on the pulse admixture proportion and admixture time, in addition to split time (Steinrücken et al., 2019). However, a single pulse admixture event and a single estimate on split time oversimplifies the complexity of human history. Especially the legacy of admixture with distinct archaic populations may affect the estimate in structured demographic models. Existing evidence estimates the introgression proportion from Neandertal to non-African populations to be around 2% (Green et al., 2010; Prüfer et al., 2014a), and the split between Neandertal and modern population occurred at around 450-550 thousands years ago (Prüfer et al., 2014a). The Denisovan gene flow is detected in Oceania at proportion 5%, particularly in Papuan and Australians (Reich et al., 2010). Moreover, several studies have indicated there was gene flow from an unknown archaic population into African populations, especially western African populations like Yoruba and Mandenka, though not convincingly shown for deeply diverged Afrian hunter gatherers San, Mbuti and Biaka (Hammer et al., 2011; Lachance et al., 2012; Lorente-Galdos et al., 2019; Plagnol & Wall, 2006). Manuscript A of this thesis proposes MSMC-IM, which is a new flexible approach, avoiding strong assumptions on the demographic model and allowing estimation of admixture with both modern-day and distinct archaic populations. It adds finer details on the global picture of population separations, particularly on the deep separations that occurred within Africa, and characterizes the post-split admixture and archaic introgression in present-day populations in a quantitative way.

1.2 Genetic perspectives on the population structure in Africa

Any newly developed analytical method of demographic reconstruction needs validation on real data - either with modern or ancient genomes. Modern human data is comparatively easy to collect and relatively cheap to sequence to high depths, therefore most approaches described above are designed for high coverage modern human genomes. To some extent, the legacy of the past human demography can be extrapolated from modern genomes with the help of carefully developed methods, and joint interpretation with archeological and linguistic assumptions. But, a much more direct way is to use aDNA, which enables us to test directly on what in the past shaped present-day genetic structure and population diversity. In this chapter I will provide an overview on the population structure of Africa derived from genetic perspectives.

1.2.1 Present-day genetic, linguistic and subsistence variations in Africa

Africa harbours the deepest genetic lineages in humans and also hosts enormous genetic, cultural and linguistic diversity. When characterizing the great genetic variations in Africa, the correlations to linguistic, cultural, and ethnic properties are of importance given over 2,000 ethno-linguistic groups have been identified there (Tishkoff et al., 2009). African languages are classified into four major macro-families: Afroasiatic, Nilo-Saharan, Niger-Congo and Khoesan (Heine & Nurse, 2000), each often linked to specific subsistence strategies. African populations practice a variety of diverse subsistence modes, including hunting-gathering, herding, farming and agro-pastoralism (i.e. mixed with farming) (Tishkoff et al., 2009). Nilo-Saharan speaking people are mainly pastoralists from central and eastern Africa, such as the Dinka, Maasai, Luo from the Nile Basin (Tishkoff et al., 2009). Afroasiatic-speaking people, of wider distribution in northern and eastern Africa, mainly practice agriculture and agro-pastoralism, such as Beja pastoralist and Oromo mixed farmers from the Horn of Africa

in the northeast (Tishkoff et al., 2009). Khoesan-speaking people, known for their unique click consonants, are hunter-gatherers indigenous to southern Africa except for the Hadza and Sandawe residing in eastern Africa. Niger-Congo is Africa's largest language family, widely distributed in western, central and southern Africa. Bantu as a subfamily, is spoken by almost half of the Niger-Congo populations. The ancient Bantu-speaking people expanded eastward and southward from their hypothesized homeland - Cameroon in west Africa, with their farming technology (de Filippo Cesare et al., 2012; Phillipson, 2005). This so-called "Bantu expansion" has transformed the local population structure in eastern Africa and southern Africa remarkably. Present-day Bantu-speaking groups in southern Africa still show genetic similarity to western African populations like Yoruba, Mandenka (Schlebusch et al., 2012).

Present-day people residing in eastern Africa, show a great diversity of regional substructure in genetics and languages (Tishkoff et al., 2009). The two indigenous click-speaking groups - Hadza and Sandawe, represent a unique eastern African hunter-gatherer lineage, reflecting the long-term presence of indigenous hunter-gatherer ancestry in eastern Africa (Pickrell et al., 2012a; Tishkoff et al., 2009). While the other groups reflect successive migration waves of Afroasiatic Cushitic-, Nilotic- and Bantu-speaking groups, who live as farmers, herders or agro-pastoralists contemporaneously nowadays (Tishkoff et al., 2009). In particular, some modern eastern and northeastern African populations show close genetic connections to Eurasia, likely driven by the Middle East and the Arab expansion followed by southward migration waves along the Nile river in Africa (Hollfelder et al., 2017; Pickrell et al., 2014; Schlebusch & Jakobsson, 2018).

East Africa was not just the destination of migration waves but also the origin of migrations. East African pastoralists brought themselves and the practice of pastoralism to southern Africa, independent of the "Bantu expansion" migration wave (Pickrell et al., 2012a, 2014; Schlebusch & Jakobsson, 2018). In southern Africa, Khoekhoe herders, indigenous San hunter-gatherers and Bantu groups composed the majority of present-day populations. Almost all modern-day Khoesan groups derive some of their ancestry from east Africa/Eurasia (Pickrell et al., 2012a) possibly as a result of the east-to-south migration of eastern African pastoralists. The indigenous San hunter-gatherer has the greatest genetic diversity and is the earliest diverged population among all modern human populations (Pickrell et al., 2012a; Schlebusch et al., 2012; Tishkoff et al., 2009).

1.2.2 Ancient DNA sheds new lights on past population movements in Africa

Genetic variation within modern African populations have distinguished several distinct genetic clusters - west African/Bantu-related groups, central African hunter-gatherers (i.e. Mbuti and Biaka), east African hunter-gatherers (i.e. Hadza and Sandawe) and south African hunter-gatherers (i.e. San), which are less admixed in comparison to other African groups. Ancient hunter-gatherer genomes, such as the 4500-year-old Ethiopian individual "Mota", 8000- to 2000-year-old Malawi individuals and 2000-year-old San, indicate an ancient east-to-south hunter-gatherer cline genetically mirroring the geography (Gallego Llorente et al., 2015; Skoglund et al., 2017), suggesting the long lasting persistence of indigenous hunter-gatherer ancestry on the Africa continent. In addition to indigenous hunter-gatherers, the spread of food producers (pastoralists and agriculturalists) has non-trivially complicated

African population structure, forming a complex mosaic of communities in different regions of Africa.

Ancient DNA from the Luxmanda site in Tanzania provides the first direct genetic evidence of the arrival of pastoralists in eastern Africa at ca. 3000 years ago (Skoglund et al., 2017). A later study proposed a multi-phased model on the formation of the eastern African pastoralist gene pool - i) during the Pastoral Neolithic, an ancestry component related to Chalcolithic Levantine groups first entered easten Africa and mixed there with local Late Stone Age foragers, and ii) during the Iron Age, herders related to Nilotic-speaking people expanded into eastern African and mixed with locals again (Prendergast et al., 2019). The prehistoric genetic connection between the Levant and northern/eastern Africa started a long time ago, since the Pleistocene in northern Africans (van de Loosdrecht et al., 2018) to 1300 years ago in ancient Egyptians (Schuenemann et al., 2017). The traces of Levantine ancestry in ancient eastern African pastoralists provides further evidence on the hypothesized continuous population movement between Levant and Africa in the past. A single ancient individual from Pemba Island documents the genetic footprint of another group of food producers - Bantu-related agriculturalists in eastern Africa (Skoglund et al., 2017). But whether the incoming food producers and indigenous hunter-gatherers mixed with each other over time remains unclear.

The admixture between food producers and indigenous hunter-gatherers has been clearly detected in southern Africa. A 1200-year-old individual from South Africa shows an apparent genetic signature of both ancient eastern African pastoralists and San hunter-gatherers (Skoglund et al., 2017), as a genetic legacy of the spread of pastoralism. Four 400-years-old South Africans are genetically close to present-day west African/Bantu-speaking populations, with little genetic contribution from indigenous southern African hunter-gatherers (Schlebusch et al., 2017). Ancient DNA allows direct measurements on the ancestral origins of individuals, disentangling their ancestry from a mosaic of pastoralist, agriculturalist and hunter-gatherer communities. In manuscript B of this thesis, I analyzed 20 newly reported ancient genomes from wide spatial and temporal space in sub-Saharan Africa, and characterize the interactions between hunter-gatherers, pastoralists and Banut-speaking groups from genetic perspectives.

1.3 A brief introduction into Eurasia's Eastern Steppe

Ancient pastoral communities are widely distributed in many parts of the world, not just in Africa but also on the Eurasia continent. Recent ancient DNA studies have identified a series of pastoralists-driven migration events on the Western Steppe transforming the regional genetic makeup during the Bronze Age in the west Eurasia (Allentoft et al., 2015; P. de B. Damgaard et al., 2018; de Barros Damgaard et al., 2018; Haak et al., 2015; Mathieson et al., 2015; C.-C. Wang et al., 2019). The Yamnaya, as the earliest representative of western steppe herders from the Pontic-Caspian steppe, has genetically contributed substantially to the people of the European Corded Ware culture at c.a. 2500BC (Haak et al., 2015), but also spread into Central Asia and the Eastern Steppe giving rise to the Afanasievo culture in Altai Mountain and Minusinsk Basin (Allentoft et al., 2015; Anthony, 2010). Later in 2000BC, the Sintashta culture emerged in the Urals with the earliest known chariots, which played an important role in goods transport, human migration and ancient warfare (Kuznetsov, 2006).

The people of the Sintashta culture show additional genetic affinity with the Corded Ware people of European Neolithic farmer ancestry, compared with the previous Yamnaya people (Allentoft et al., 2015). The Sintashta forms part of the Bronze Age Andronovo horizon, and is genetically indistinguishable from people from the core area of the Andronovo culture (Allentoft et al., 2015). Technological innovation of transport brings the Sintashta's metal to the Bactria–Margiana Archaeological Complex in central Asia (Anthony, 2010). To date, the genetic evidence has shown the expansion of western Steppe herders reached as far as Central Asia, South Asia and even the periphery of the Eastern Steppe zone in Altai during the Bronze Age (Allentoft et al., 2015; P. de B. Damgaard et al., 2018; de Barros Damgaard et al., 2018; Narasimhan et al., 2019). The Eastern Steppe zone is a vast expanse of grasslands, forest-steppe and desert-steppe centered in present-day Mongolia while covering parts of modern-day China and Russia. The ecological environment makes it perfect for pastoralism. However it is still not well understood when and where the western Steppe herders came into contact with local people that inhabited the Eastern Steppe, and how pastoralists became dominant populations over thousands of years in the Eastern Steppe.

Before the Bronze Age, the Eastern Steppe zone was populated mainly by hunter-gatherers and fishers in waterside regions like Lake Baikal. Recent paleogenomic studies have revealed a strong west-east admixture cline of ancestry stretching from the Botai culture in central Kazakhstan to Lake Baikal in southern Siberia to Devil's Gate Cave in the Russian Far East (de Barros Damgaard et al., 2018; Jeong et al., 2018; Sikora et al., 2019; Siska et al., 2017). From the early Bronze Age, pastoralism was introduced in multiple phases to the Eastern Steppe, which drastically changed local lifeways and subsistence strategies (Honeychurch, 2015; Kindstedt & Ser-Od, 2019). The first migration wave of Steppe herders to the east is represented by the Afanasievo (3000 BCE), centered in the Upper Yenisei region, who are genetically indistinguishable from the Yamnaya in Pontic-Caspian area (Allentoft et al., 2015). The later culture called Chemurchek (2750-1900 BCE), centered in the southern Altai-Syan regions, is of controversial origin despite the clear cultural influence from Afansievo (Kovalev, 2014). Whether pastoralists left a genetic legacy in the Chemurchek people is questionable. By the Middle/Late Bronze Age (MLBA, 1900-900 BCE), pastoralists' ruminant dairying became prevalent in western and northern Mongolia at sites associated with the Deer Stone-Khirigsuur Complex (DSKC), and in eastern Mongolia associated with the Ulaanzuukh culture (Jeong et al., 2018). This raised three questions on the hypothesized second phase of pastoralism introduction and spread: i) how long did Afanasievo-related pastoralists survive in the Eastern Steppe; ii) is the widespread dairying during the MLBA associated with a new source of western Steppe herders different from the Afanasievo; iii) if there was a new migration wave from the west, how far east did they spread during the MLBA.

A recent study by Jeong et al 2018 (Jeong et al., 2018) addressed low genetic contribution from western Steppe herders represented by the Sintashata into DSKC-related individuals from northern Mongolia, suggesting the appearance of western Steppe ancestry from a new source (although in low proportion during MLBA). However, the broader picture is still lacking. The associations between the DSKC and Ulaanzuukh groups remain poorly understood, and little is known about other MLBA burial traditions in Mongolia such as the Mönkhkhairkhan and Baitag. By the end of the second millennium BCE, the mainstream culture in the Eastern Steppe started shifting from previously MLBA cultures to the Early Iron

Age culture - Slab Grave (ca. 1000-300 BCE). Meanwhile, the western periphery of Mongolia and eastern Kazakhstan became the direct contact zone with Iron Age Scythian herders who widely flourished across the entire Eurasian Steppe. The Uyuk culture (ca. 700-200 BCE) from the Sayan mountains, also known as the Aldy-Bel culture, had strong cultural links to the Scythian nomads of the Altai - the Pazyryk (ca. 500-200 BCE) and Saka (ca. 900-200 BCE) cultures (Savinov, 2002; Tseveendorj, 2007). It is unclear whether the cultural links from the archaeological perspective still stand from a genetic perspective.

From the first millennium, the Eastern Steppe is the political center of many pastoral nomadic empires, notably the Xiongnu (209 BCE-98 CE), and the Mongol (1206-1368 CE). The Mongol empire, known for its founder Genghis Khan, was the largest contiguous empire that eventually stretched from eastern Europe to the Sea of Japan. The high mobility of these nomadic people raises many unanswered questions, such as whether the formation of these nomadic empires relates to the preceding prehistoric cultures, how the genetic transition between consecutive empires was, and what is the impact of these ancient historic polities on present-day Mongolian genetic diversity. In manuscript C of this thesis, I analyzed genome-wide data of newly reported 214 individuals from Mongolia and Russia spanning nearly 6,000 years (ca. 4600BCE to 1400CE). I present a broad picture of population movements and dynamic population histories by characterizing major genetic clusters from different time layers, reconstructing the contact history between the east and west over time, and illustrating the temporal and spatial changes of the gene pool in the Eastern Steppe.

2. Aim of the thesis

The aim of this thesis is to reconstruct human population structure with genomic data through: i) developing a new method of estimating time-dependent migration rate in deep time depth; ii) analyzing ancient DNA from archaeological human remains for continental-wise research questions. Since the origin of modern humans from Africa, the dynamics of population separations, movements and admixture over time have shaped the complex population structure in present-day human populations. Characterizing these demographic features in a continuous manner is currently lacking in the field. Many demography-inference methods utilize strictly structured population-split models using African indigenous hunter-gatherer groups like San or Mbuti as the early diverged distant group. Despite the long history of human habitation in Africa, genomic studies on African populations have been underrepresented in a long time. Noticeably, the migration of food producers - pastoralists and agriculturalists, according to archaeological and linguistic studies, has restructured the co-existing composition of local communities in Africa to a great extent. A similar demographic change has been observed in the late Neolithic Europe where pastoralists from Pontic-Caspian and farmers from the Middle East transformed the previous hunter-gatherer-dominant population structure meanwhile bringing in their techniques. A more extreme case is in the Eastern Steppe where modern populations still maintain pastoralism as their main subsistence nowadays. There has been great focus on studying population movements and admixture in the Western Eurasia, while the past of population migration and settlements in the Eastern Steppe is largely unknown.

To reconstruct human demography history, this thesis presents a new analytical method with application to modern-day worldwide whole genome data, and addresses main demographic events in Africa and the Eastern Steppe zone of Eurasia by analyzing ancient DNA.

Manuscript A:

- How did human population structure develop, and since when?
- How populations separated through periods of isolation, partial gene flow and admixture over time?
- How to characterize population separations continuously and quantitatively?

Manuscript B:

- What is the genetic association between ancient hunter-gatherers from eastern Africa and from other regions of Africa?
- How was the spread of pastoralists northern to eastern Africa during the Pastoral Neolithic period and how they interacted with hunter-gatherers on the genomic level?
- How agropastoralists of Nilotic-related expansion and Bantu-related migrations changed gene pool in Eastern Africa during the Iron Age?
- What is the genetic impact of the arrival of eastern African pastoralists and Bantu farmers in southern Africa? Who arrived in southern Africa first?

Manuscript C:

- What was the temporal and spatial structure of genetic profile in the Eastern Steppe in the past?

- How does genetic structure correlate with prehistoric cultural labels and historical records?
- What was the genetic picture of the Eastern Steppe before and after the introduction of pastoralism, and to what extent incoming pastoralists transformed local population structure?
- How was the formation of pastoral nomadic empires genetically?
- Does dairy pastoralism have an impact on the local gene pool, such as adaptive alleles about lactase persistence?

3. Overview of Manuscripts and author's Contribution

3.1 Manuscript A

"Tracking human population structure through time from whole genome sequences"

Ke Wang, Iain mathieson, Jared O'Connell, Stephan Schiffels

Published at PLOS Genetics (March 2020)

In Manuscript A, we present a novel approach based on the Multiple Sequentially Markovian Coalescent (MSMC) to analyze the population separation history in a continuously parameterized fashion. Our new method called MSMC-IM, quantifies the population separations and migrations through a piecewise constant migration rate across populations, which provides a direct time-dependent estimate of gene flow for the first time.

MSMC (Schiffels & Durbin, 2014b) introduced the concept of the relative cross coalescence rate for characterizing the separation process continuously without the specification of an explicit demography model. The relative CCR allows a hypothesis-free manner to estimate key aspects of population separation, such as using the time at which lineages are half as likely to coalesce between rather than within populations as a heuristic estimate for the split time. However interpreting the relative CCR without any explicit model in more complex demographic scenarios is challenging. Our new approach MSMC-IM, overcomes this disadvantage by fitting a continuous IM model to the distribution of coalescence times estimated from MSMC's piecewise constant model, while maintaining the flexibility character on result-interpreting.

We show that MSMC-IM can identify multiple demographic scenarios from clean-split separation to separations with post-split admixture and even archaic introgression. By applying MSMC-IM to worldwide human genomic data from 15 populations, we track the process of human genetic diversification via the estimated time-dependent migration rates across pairs of populations. We obtain a global picture of human separation and migration history from a million years ago to recent thousand years ago. In particular, we detect traces of extremely deep ancestry between some African populations, with around 1% of ancestry dating to population divergence older than a million years ago.

Author's contributions:

Stephan Schiffels initiated, designed and supervised this study. Stephan Schiffels and I conceptualized the model, and I implemented and tested it on simulated genetic data. I generated and processed simulation data and processed human genetic data from published datasets from Prüfer et al 2014. Iain Mathieson processed modern human data from the Simons Genome Diversity Project dataset (Mallick et al. 2016) and carried out the test on comparing different phasing strategies. Jared O'Connell generated and provided the high coverage genome of an aboriginal Australian sequenced with DNA libraries of long reads. I analysed the data from simulations from public datasets, and interpreted results together with Stephan Schiffels. I wrote the majority of the paper with considerable contributions from Stephan Schiffels and input all other co-authors. During the review process, I improved the model design and added more tests from simulations as reviewers requested.

Model development	50%
Method implementation	100%
Model testing	100%
Human genetic data analysis	80%
Manuscript writing	60%

3.2 Manuscript B

"Ancient genomes reveal complex patterns of population movement, interaction and replacement in sub-Saharan Africa"

Ke Wang[†], Steven Goldstein[†], Madeleine Bleasdale, Bernard Clist, Koen Bostoen, Paul Bakwa-Lufu, Laura T. Buck, Alison Crowther, Alioune Dème, Roderick J. McIntosh, Julio Mercader, Christine Ogola, Robert C. Power, Elizabeth Sawchuk, Peter Robertshaw, Edwin N. Wilmsen, Michael Petraglia, Emmanuel Ndiema, Fredrick K. Manthi, Johannes Krause, Patrick Roberts, Nicole Boivin and Stephan Schiffels

[†]equal contributors

Published at Science Advances (June 2020)

In Manuscript B, we report genome-wide data of twenty sub-Saharan African genomes ranging from four thousand years ago to recent hundreds of years, including the first ancient DNA from the Demographic Republic of the Congo, Uganda and Botswana. The high genetic heterogeneity in present-day Africa is shaped by complex patterns of population movement, interaction and replacement. Particularly, the spread of food producers, such as the big migration of Bantu-speaking farmers and the dispersal of pastoralists from eastern to southern Africa, has a great impact on the indigegnous hunter-gatherer communities. To enlighten the early population movements and admixture in Africa, we sampled ancient individuals from the key regions of significant interaction between farmers, pastoralists and local hunter-gatherers, especially from eastern and southern Africa.

We find eastern Africa witnessed population-level interactions millennia ago between local eastern Africa foragers and groups whose forager descendants are today restricted to central and southern Africa. We interpret this phenomenon as the potential contraction of once overlapping hunter-gatherer ancestries in eastern Africa. We also record the formation of pastoralists' genetic makeup in eastern Africa occurring at two time frameworks - one related to incoming Levantine ancestry during Pastoral Neolithic and the other associated with Nilotic expansion during Iron Age Pastoral period. Noticeably, the admixture between pastoralists and foragers show a more complex pattern without following a specific chronological order, which suggest, in eastern Africa, communities with high or unadmixed hunter-gatherer-related ancestry continued to live alongside communities with high or unadmixed Pastoral-Neolithic related ancestry until nearly the Iron Age.

We, using ancient genomes, directly document the arrival of pastoralist-related ancestry at the first millenium in northeastern Congo and Botswana, and show that admixture between pastoralists and foragers precedes incorporation of Bantu ancestry in Iron Age northern Botswana, suggesting pastoralism arrives southern Africa earlier than farming. Historical individuals from the west coastline of the DR Congo record the genetic footprint of Bantu-speaking people in an un-admixed format hundred years ago. Overall, our genetic findings, together with archeological information, highlight how migration and admixture have dramatically reshaped the genetic map of sub-Saharan Africa in the last few millennia.

Author's contributions:

- These authors contributed equally to this work: **Ke Wang**, Steven Goldstein
- These authors jointly supervised this work: Nicole Boivin, Stephan Schiffels

Stephan Schiffels and Nicole Boivin initiated, designed, and supervised this study. Steven Goldstein, Madeleine Bleasdale, Bernard Clist, Koen Bostoen, Paul Bakwa-lufu, Laura T. Buck, Alison Crowther, Alioune Dème, Roderick J. McIntosh, Julio Mercader, Christine Ogola, Robert C. Power, Elizabeth Sawchuk, Peter Robertshaw, Edwin N. Wilmsen, Michael Petraglia, Emmanuel Ndiema, Fredrick K. Manthi, Patrick Roberts, Nicole Boivin either collected archaeological materials or helped on providing the relevant archaeological and historical context of samples. The technical laboratory staff carried out all required laboratory procedures for ancient DNA analysis, including sampling of archaeological material, DNA extraction and library preparation. Johannes Krause and Stephan Schiffels supervised the laboratory work and sequencing.

I did bioinformatical processing on whole-genome sequencing data from 57 ancient sub-Saharan African individuals and selected 20 genomes that yielded enough amount of authentic ancient human DNA for in-depth sequencing following in-solution SNP capture. For the capture data, I conducted quality control on the whole genome and determined sex with X- and Y-chromosomal DNA. I performed population genetic analyses using autosomal DNA, and assigned the haplogroup of mitochondrial and Y-chromosomal DNA. I interpreted genetic results in the archaeological context with Stephan Schiffels, Steven Goldstein and Nicole Boivin. I prepared all figures and tables required for the paper, and wrote the majority of the paper (except the archeological context in the Supplementary Text) with Steven Goldstein, Nicole Boivin, Stephan Schiffels and input from all other coauthors. During the review process, I conducted additional analyses and rephrased some paragraphs as reviewers requested, with input from Steven Goldstein, Nicole Boivin, Stephan Schiffels and all other coauthors.

Sample procurement	0%
Laboratory work	0%
Bioinformatic data processing	100%
Population genetic analysis	100%
Manuscript writing	60%

3.3 Manuscript C

"A dynamic 6,000-year genetic history of Eurasia's Eastern Steppe"

Choongwon Jeong[†], Ke Wang[†], Shevan Wilkin, William Timothy Treal Taylor, Bryan K. Miller, Sodnom Ulziibayar, Raphaela Stahl, Chelsea Chiovelli, Jan H. Bemmann, Florian Knolle, Nikolay Kradin, Bilikto A. Bazarov, Denis A. Miyagashev, Prokopiy B. Konovalov, Elena Zhambaltarova, Alicia Ventresca Miller, Wolfgang Haak, Stephan Schiffels, Johannes Krause, Nicole Boivin, Erdene Myagmar, Jessica Hendy, Christina Warinner

[†]equal contributors

Published at Cell (November 2020)

In Manuscript C, we report genome-wide data of 214 ancient individuals in the Eastern Steppe of Eurasia continent spanning 6,000 years from ca. 4600BCE to 1400CE. In this largest east Asian genomics study to date, we report for the first time, the dynamic changes of population history in the Eastern Steppe including demographic events associated with subsistence changes and the formation of the nomad pastoral empire.

The Eastern Steppe was sparsely populated by hunter-gatherers since the mid-Holocene, who left long-lasting genetic footprint in present-day Tugngusic- and Nivkh-speaking populations of the Far East. We refer to this profile as “Ancient Northeast Asian” (ANA) in contrast to another widespread mid-Holocene genetic profile known as “Ancient North Euraisian” (ANE) from Siberia. Six pre-Bronze Age individuals in our study highlight the broader genetic shift in hunter-gatherers with slightly inflating ancestry of ANE from early Neolithic to Bronze Age.

We demonstrate the eastward migration of western steppe herders (Afanasievo) during the Early Bronze Age extended around 1500km further east than previously shown, reaching and introducing dairy pastoralism to central Mongolia by 3000BCE. We also find the subsequent Early Bronze Age Chemurchek culture in the Mongolian Altai did not derive their ancestry from earlier Afanasievo migrants who both share close cultural features, but rather represents an independent migration. During the Middle and Late Bronze Age (MLBA), we identify a strong geographic-genetic structure of three distinct gene pools in west, north and south-central Mongolia, with west Mongolia as the frontier of encountering the MLBA western herders (the Sintasha). The spatial structure still stands during the Early Iron Age, but breaks down afterwards when the first nomadic pastoral empire - the Xiongnu arose.

We find the formation of the Xiongnu was associated with the mixture of these previously separated populations and a rapid influx of new gene flows from surrounding regions. The later empires in early Medieval, such as Turkic, Uigur, show high genetic heterogeneity tracing their ancestry to various regions across Eurasia. Until Mongol empire - the largest historical pastoral empire to date, Mongol-period individuals show a remarkable increase in eastern Eurasian ancestry, and they mark the first appearance of a genetic profile that resembles present-day Mongolic-speaking populations habitating Mongolia and the surrounding area. In particular, despite the deep history of dairy pastoralism in the Eastern Steppe (>5000 years), we find low frequency of adaptive alleles related to lactase persistence (LP) over the entire period and showed no evidence of selection, raising the question on the role of LP in prehistoric dairying. All in all, our study illuminates previously uncharacterized patterns of complex interplay between genetic, sociopolitical, and cultural changes on the Eastern Steppe.

Author’s contributions:

- These authors contributed equally to this work: Choongwon Jeong, **Ke Wang**
- These authors jointly supervised this work: Choongwon Jeong, Christina Warinner

Christina Warinner, Choongwon Jeong, Erdene Myagmar, Nicole Boivin initiated and designed this study. Christina Warinner and Choongwon Jeong supervised the study. Erdene Myagmar, Shevan Wilkin, Jessica Hendy, Jan Bemmann, Sodnom Ulziibayar, Wolfgan Haak, Florian Knolle, Bilikto A. Bazarov, Denis A. Miyagashev, Prokopiy B. Konovalov, Elena Zhambaltarova, Alicia Ventresca Miller, Nicole Boivin, and Christina

Warinner provided archeological materials and resources. Rapheala Stahl did all required ancient DNA laboratory work, with supervision from Choongwon Jeong. Choongwon Jeong processed raw genetic data, and performed quality control, kinship analyses, mitochondrial and Y-chromosomal haplogroup assignments of the majority of the dataset. Florian Knolle and Wolfgang Haak conducted these procedures for 13 individuals from Baikal region.

I got involved into the project when the majority of the dataset was ready, and performed bioinformatic data processing for leftover individulas including merging sequencing data from different DNA libraries of the same individual. I carried out all population genetic analyses for the complete dataset of 214 invididuals, including analyses of populaton structure and relationships, admixture modelling, dating admixture events, phenotypic SNP analysese and investigation of sex-biased admixture. Bryan Miller, William Taylor, Jan Bammann, Erdene Myagamar provided the archaeological context of these inviduals. I integrated genetic results with the archaeological background together with Bryan Miller, William Taylor, Choongwon Jeong and Christina Warinner. Ke Wang, Choongwon and Christina Warinner wrote the paper, with contributions from Bryan Miller, William Taylor, Jan Bammann and all other coauthors. I prepared all figures with genetic results and tables required for the paper, and contributed most to the text related to the genetic analyses during the preparation of the manuscript and during the revision process.

Sample procurement	0%
Laboratory work	0%
Bioinformatic data processing	20%
Population genetic analysis	100%
Manuscript writing	60%

4. Manuscript A

PLOS GENETICS

RESEARCH ARTICLE

Tracking human population structure through time from whole genome sequences

Ke Wang[1], **Iain Mathieson**[2], **Jared O'Connell**[3], **Stephan Schiffels**[1]*

1 Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany, **2** Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **3** 23andMe Inc., Mountain View, California, United States of America

* schiffels@shh.mpg.de

Citation: Wang K, Mathieson I, O'Connell J, Schiffels S (2020) Tracking human population structure through time from whole genome sequences. PLoS Genet 16(3): e1008552. https://doi.org/10.1371/journal.pgen.1008552

Editor: Mikkel H. Schierup, Aarhus University, DENMARK

Received: March 25, 2019

Accepted: December 4, 2019

Published: March 9, 2020

Data Availability Statement: SGDP data (Mallick et al. 2016) are available from https://reichdata.hms.harvard.edu/pub/datasets/sgdp/. Present-day human genome sequences published in Prüfer et al. 2014 (Nature) are available from http://cdna.eva.mpg.de/neandertal/altai/ModernHumans/.

Funding: SS and KW acknowledge support by the Max Planck Society. IM was supported by a Research Fellowship from the Alfred P. Sloan foundation [FG-2018-10647] and a New Investigator Research Grant from the Charles E. Kaufman Foundation [KA2018-98559]. The

Abstract

The genetic diversity of humans, like many species, has been shaped by a complex pattern of population separations followed by isolation and subsequent admixture. This pattern, reaching at least as far back as the appearance of our species in the paleontological record, has left its traces in our genomes. Reconstructing a population's history from these traces is a challenging problem. Here we present a novel approach based on the Multiple Sequentially Markovian Coalescent (MSMC) to analyze the separation history between populations. Our approach, called MSMC-IM, uses an improved implementation of the MSMC (MSMC2) to estimate coalescence rates within and across pairs of populations, and then fits a continuous Isolation-Migration model to these rates to obtain a time-dependent estimate of gene flow. We show, using simulations, that our method can identify complex demographic scenarios involving post-split admixture or archaic introgression. We apply MSMC-IM to whole genome sequences from 15 worldwide populations, tracking the process of human genetic diversification. We detect traces of extremely deep ancestry between some African populations, with around 1% of ancestry dating to divergences older than a million years ago.

Author summary

Human demographic history is reflected in specific patterns of shared mutations between the genomes from different populations. Here we aim to unravel this pattern to infer population structure through time with a new approach, called MSMC-IM. Based on estimates of coalescence rates within and across populations, MSMC-IM fits a time-dependent migration model to the pairwise rate of coalescences. We implemented this approach as an extension to existing software (MSMC2), and tested it with simulations exhibiting different histories of admixture and gene flow. We then applied it to the genomes from 15 worldwide populations to reveal their pairwise separation history ranging from a few thousand up to several million years ago. Among other results, we find evidence for remarkably deep population structure in some African population pairs, suggesting that deep ancestry dating to one million years ago and older is still present in human populations in small amounts today.

funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: Jared O'Connell is employed by 23andMe Inc. The authors have declared that no competing interests exist.

Introduction

Genomes harbor rich information about population history, encoded in patterns of mutations and recombinations. Extracting that information is challenging, since in principle it requires reconstructing thousands of gene genealogies separated by ancestral recombination events, using only the observable pattern of shared and private mutations along multiple sequences. One important innovation was the Sequentially Markovian Coalescent (SMC) model [1,2], which is an approximate form of the ancestral recombination graph that can be fitted as a Hidden Markov model along the sequence. This approach has been used to infer demographic history in methods like PSMC [3], MSMC [4], diCal [5,6] and SMC++ [7].

These methods estimate one or both of two important aspects of population history: i) The history of the effective population size, and ii) the history of population structure. The second aspect, which entails reconstructing the timing and dynamics of population separation requires a non-trivial choice of parameterization: While methods like diCal2 [5], as well as many methods based on the joint site frequency spectrum [8–11] use an explicit population model with split times, migration rates or admixture events, MSMC [4] introduced the concept of the relative cross coalescence rate to capture population separations in a continuously parameterized fashion. The main advantage of that approach is that it does not require the specification of an explicit model, but can be applied hypothesis-free to estimate key aspects of population separation, for example the time at which lineages are half as likely to coalesce between rather than within populations, which is often used as a heuristic estimate for the divergence time between the populations. A disadvantage is that other important aspects of population separation, like post-split or archaic admixture, are non-trivially encoded in features of the cross-coalescence rate other than this mid-point. As a consequence, it is difficult to interpret the cross-coalescence rate in terms of actual historical events.

Here, we propose an approach to overcome the disadvantages of the relative cross coalescence rate, while maintaining the continuous character of population separation from MSMC without explicitly specifying a complex population phylogeny. We present a new method MSMC-IM, which fits a continuous Isolation-Migration (IM) model to the distribution of coalescence times, estimated from MSMC's piecewise constant model. In MSMC-IM, separation and migration between a pair of populations is quantified by a piecewise constant migration rate across populations, and piecewise constant population size changes within each population. We apply our method on world-wide human genomic data from the Simons Genome Diversity Project (SGDP) [12] to investigate the history of global human population structure.

Results

Estimating pairwise coalescence rates with MSMC2 and fitting an IM model

To model the ancestral relationship between a pair of populations, we developed an isolation-migration model with a time-dependent migration rate between a pair of populations, which we call MSMC-IM. The approach requires time-dependent estimates of pairwise coalescence rates within and across two populations. To estimate these rates, we use an extension of MSMC [4], called MSMC2, which was first introduced in Malaspinas et al. 2016 [13] (Fig 1A, Methods). MSMC2 offers two key advantages over MSMC [4]. First, the pairwise coalescence model in MSMC2 is exact within the SMC' framework [2], whereas MSMC's model uses approximations that cause biases in rate estimates for larger number of haplotypes (S1 Fig). Second, since MSMC2 uses the pairwise tMRCA distribution instead of the first tMRCA distribution, it estimates coalescence rates within the entire range of coalescence events between

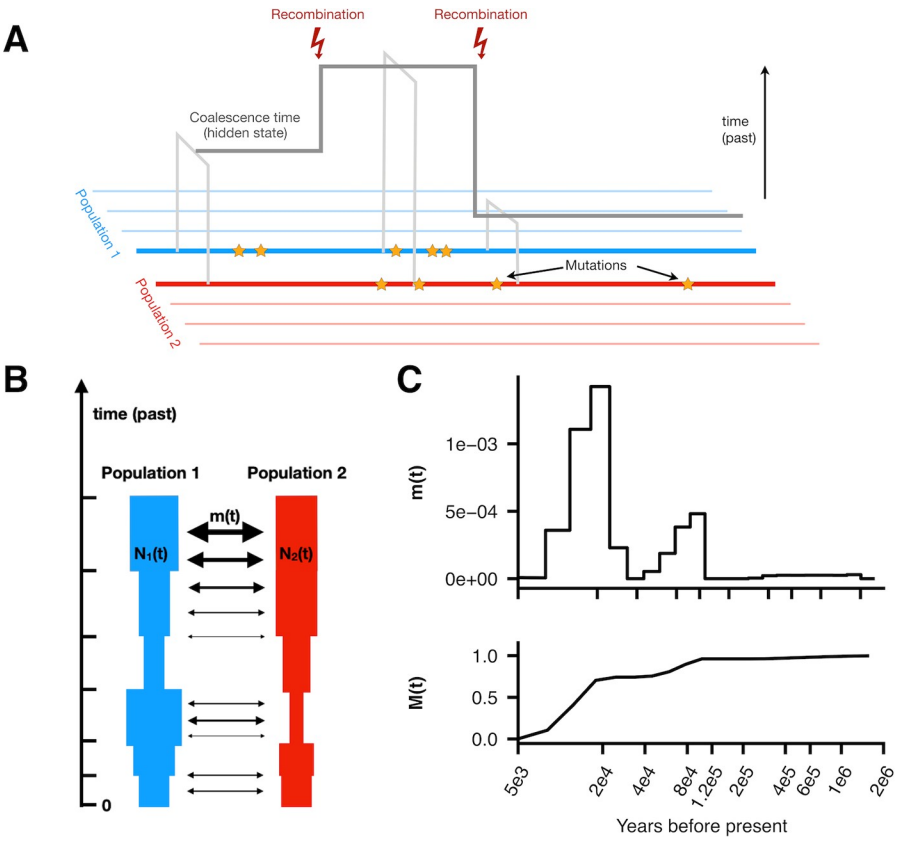


Fig 1. Schematic of MSMC2 and MSMC-IM. (A) MSMC2 analyses patterns of mutations between pairs of haplotypes to estimate local coalescence times along the genome. (B) MSMC-IM fits an isolation-migration model to the pairwise coalescence rate estimates, with time-dependent population sizes and migration rate. (C) As a result, we obtain the migration rate over time, $m(t)$, and the cumulative migration probability, $M(t)$, which denotes the probability for lineages to have merged by the time t and which we use to estimate fractions of ancestry contributed by lineages diverged deeper than time t.

https://doi.org/10.1371/journal.pgen.1008552.g001

multiple haplotypes, which ultimately increases resolution not just in recent times but also in the deep past. These two improvements are crucial for our new method MSMC-IM, which relies on unbiased coalescence rate estimates within and across populations, in particular in the deep past. Specifically, MSMC2 recovers simulated population size histories (with human-like parameters) well up to 3 million years ago, while keeping the same high resolution in recent times as MSMC (S1 Fig).

Given MSMC2's estimates of time-dependent coalescence rates within populations, $\lambda_{11}(t)$ and $\lambda_{22}(t)$, and across populations, $\lambda_{12}(t)$, we use MSMC-IM to fit an Isolation-Migration (IM) model to those three coalescence rates (see Methods). MSMC-IM's model assumes two populations, each with its own population size $N_1(t)$ and $N_2(t)$, and a piecewise-constant symmetric migration rate $m(t)$ between the two populations (Fig 1B, see Methods and S1 Text for details). Expressing the separation history between two populations in terms of a variable migration rate instead of the more heuristic relative cross coalescence rate facilitates interpretation, while maintaining the freedom to analyze data without having to specify an explicit model of splits and subsequent gene flow. Of the new parameters, the time-dependent migration rate $m(t)$ is arguably the most interesting one, and it can be visualized in two ways (Fig 1C). First, the rates themselves through time visualize the timing and dynamics of separation processes, and

second, the cumulative migration probability $M(t)$ defined as

$$M(t) = 1 - exp(-\int_0^t m(t')dt')$$

which can be understood as the proportion of ancestry that has already merged at time t, and which makes it possible to quantify proportions of gene flow or archaic ancestry through time, as illustrated below. Being by definition monotonically increasing and bounded between 0 and 1, $M(t)$ also turns out to be numerically close to the relative cross coalescence rate from MSMC [4]. When $M(t)$ becomes very close to 1, it means that lineages between the two extant populations have completely mixed into essentially one population. As a technical caveat, this means that at that time point our three-parameter model is overspecified. To avoid overfitting, we therefore employ regularization on $m(t)$ and the difference of the two population sizes (see Methods).

Evaluating MSMC-IM with simulated data

We illustrate MSMC-IM by applying it to several series of simulated scenarios of population separation (see Methods). First, the *clean-split* scenario consists of an ancestral population that splits into two subpopulations at time T (Fig 2A). Second, the *split-with-migration* scenario adds an additional phase of bidirectional gene flow between the populations after they have split (Fig 2B). Third, the *split-with-archaic-admixture* scenario involves no post-split gene flow, but contains additional admixture into one of the two extant populations from an unsampled "ghost" population, which splits from the ancestral population (Fig 2C) at time $T_a > T$. In addition, to understand how MSMC-IM behaves under asymmetric demographic histories in the two populations, we consider the *archaic-admixture-with-bottleneck*-scenario (see Fig 2D). For each scenario, we simulated 8 haplotypes (four from each population), used human-like evolutionary parameters and varied one key parameter to create a series of related scenarios (see Methods). As discussed further below, to test internal consistency, we confirmed that MSMC-IM is able to infer back its own model, using simulations based on some of the genomic inferences carried out below.

In the *clean-split* scenario, we find that MSMC-IM's inferred migration rate $m(t)$ displays a single pulse of migration around the simulated split time T (Fig 2A). This is expected, since in our parametrization, a population split corresponds to an instantaneous migration of lineages into one population at time T, thereby resulting in a single pulse of migration. In the *split-with-migration* series, we expect two instead of one pulse of migration: one at time T, as above, and a second more recent one around the time of post-split migration. In cases where the split time and migration phase are separated by more than around 20,000 years, this is indeed what we see (Fig 2B), although with some noise around this basic pattern. For less time of separation of the two migration pulses, MSMC-IM is not able to separate them in this scenario.

We also find two phases of migration for the *split-with-archaic-admixture* scenario, but this time with one phase around time T, and another one around the time of divergence of the archaic population T_a (Fig 2C). To understand this, consider how lineages in the two extant populations merge into each other (Fig 3B). One fraction $1\text{-}\alpha$ will merge into each other at the population split time T, as in the *clean-split* scenario. The other fraction, α, will merge back only at the deep divergence time of the archaic lineage. These two merge events correspond to the two pulses we observe in Fig 2C—one at T and the other at the divergence time with the archaic population, T_a. Note that unlike in the above *split-with-migration* case, here there is no signal at the time of introgression, but only at the two split times. Inferring these two migration pulses in the presence of archaic admixture is robust to demographic events, as we show with

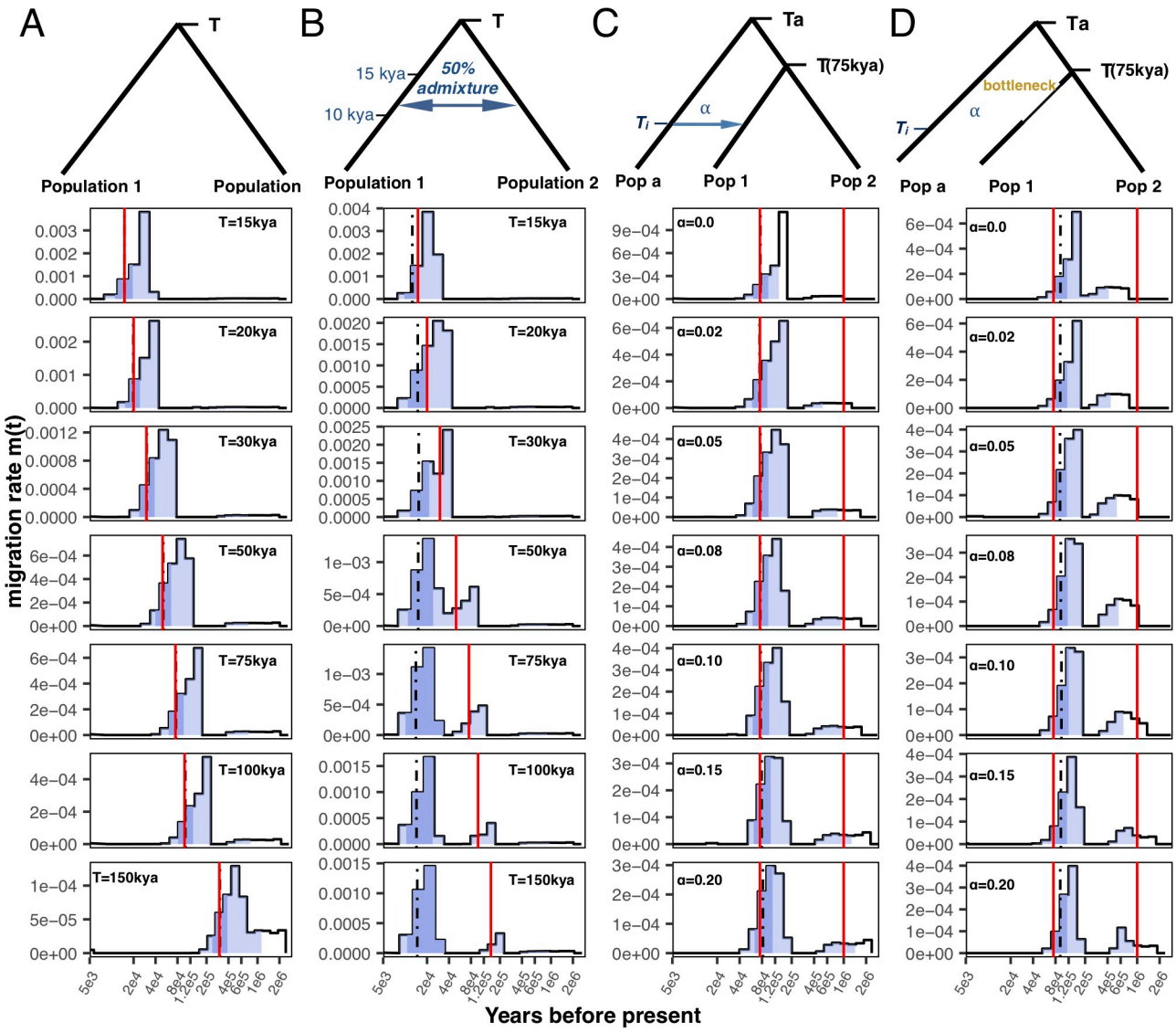


Fig 2. Simulation results. (A) *Clean-split* scenario: Two populations with constant size 20,000 each diverged at split time T in the past, varying from 15kya to 150kya. (B) *Split-with-migration* scenario. Similar to A), with T varying between 15-150kya, and a post-split time period of symmetric migration (amounting to a total migration rate of 0.5 in both directions) between 10 and 15kya. (C) *Split-with-archaic-admixture* scenario: Similar to A), with T = 75kya, and population 1 receiving an admixture pulse at 30kya from an unsampled population that separates from the ancestral population at 1 million years ago. The admixture rate varies from 0% to 20%. (D) *Split-with-archaic-admixture&bottleneck* scenario: Similar to C), but with an added population bottleneck with factor 30 in population 1 between 40-60kya. Solid red lines indicate split times in all panels. In all plots, the blue light blue shading indicates the interval between 1–99% of the cumulative migration probability, the dark blue shading from 25–75%, and the black dashed vertical line indicates the median.

https://doi.org/10.1371/journal.pgen.1008552.g002

the *archaic-admixture-with-bottleneck* scenario (Fig 2D), in which we introduced a bottleneck in one of the two extant population branches, similar in strength to the one observed in Non-African populations around 60 thousand years ago (kya) [4]. We find, however, that in the presence of a bottleneck the second pulse is a bit more recent than expected (here at 1 million years ago).

We can analyze these multiple phases of migration in a more quantitative way, by using the cumulative migration probability, $M(t)$, as introduced above. $M(t)$ monotonically increases from 0 to 1 in all scenarios, exhibiting plateaus with gradient zero at times of no migration,

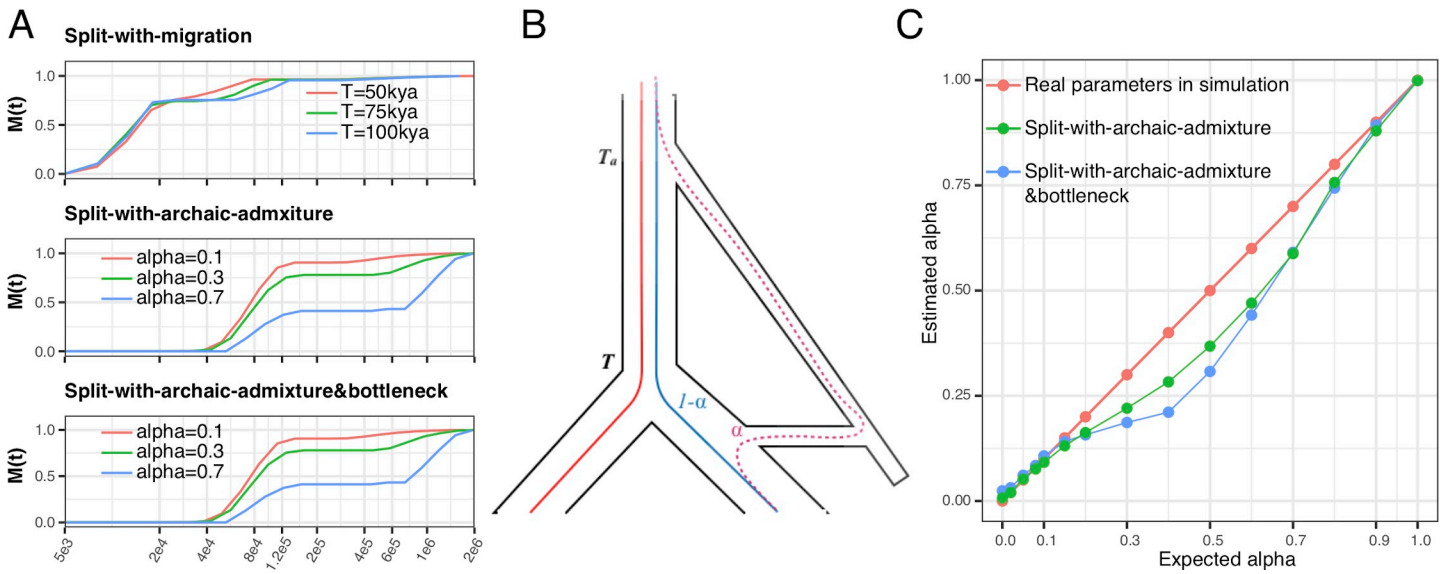


Fig 3. Evaluating admixture proportions through *M(t)*. (A) The cumulative migration probability *M(t)* is shown for selected simulation scenarios described in Fig 2B, 2C and 2D. Plateaus of *M(t)* indicate periods of isolation, with the level of the plateau indicating how much ancestry has merged before. (B) Schematic coalescence in the *Split-with-archaic-admixture* scenario. In this scenario, a fraction *1-α* of lineages sampled from the two extant populations merges at time *T*, and the rest, of proportion *α* merges as time T_a. (C) For the *split-with-archaic-admixture* scenarios (with and without bottleneck), we can use the level of the plateau in *M(t)* to estimate *1-α*, and thus *α*. The level of the plateau is measured at time *t* = 300kya.

https://doi.org/10.1371/journal.pgen.1008552.g003

and positive gradients in periods of migration (Fig 3A and S2 Fig). The level of these plateaus is indicative of how much ancestry has already merged at this point in time. Consider first the *split-with-migration* series (Fig 3A top panel), for which *M(t)* exhibits a plateau between the two migration pulses, at a level that corresponds to the amount of ancestry that has merged through the migration event. For this scenario, based on the simulated post-split migration rate between the two populations, we expect this plateau to be at around 0.64 (following the calculation in formula (64) in S1 Text). We find it to be higher than that, around 0.75, which we discuss further below. Consider now a scenario with archaic admixture (Figs 2C, 2D and 3A middle and bottom panels). At time *T*, at which both extant populations merge into each other, the cumulative migration probability reaches a plateau at a level around *1-α*, reflecting the fact that a proportion *α* has not yet merged at point *T*, but is separated by a deeply diverged population branch. Only at time T_a, this branch itself merges into the trunk of the extant populations, thereby increasing *M(t)* from *1-α* all the way to 1. Based on this rationale, we can use visible plateaus in *M(t)* to estimate fractions of archaic or otherwise deep ancestry. Indeed, this rationale leads to estimates of archaic admixture proportions in our simulations which are accurate and robust to bottlenecks for rates of *α* up to about 20%. For larger introgression rates, we find our estimates to be slightly underestimated. We attribute this to MSMC's tendency to "overshoot" changes in coalescence rates, as can be seen in the relative cross-coalescence rates for larger values of alpha (S2C and S2D Fig), which causes the level of the plateau in *M(t)* to be higher than *1-α*, and hence α to be underestimated. This is also the reason for the above-mentioned overestimation of the plateau in the *split-with-migration* scenario (Fig 3A top panel). This effect is more severe in the presence of a bottleneck (Fig 3C, blue curve) than without a bottleneck. Importantly, though, we find no evidence that *M(t)* exhibits plateaus below 1 in the absence of true deep ancestry, so this method can be considered conservative for detecting deep ancestry.

MSMC-IM also fits population sizes, which can be compared to the raw estimates from MSMC, i.e. to the inverse coalescence rates within population 1 and 2, respectively (see S1 Text for some non-trivial details on this comparison). We find that estimates for $N_1(t)$ and $N_2(t)$ are in fact close to the inverse coalescence rates, with some deviations seen in deep times, and in cases of archaic admixture. The latter is expected, given that estimated coalescence rates from MSMC2 capture both population size changes and migration processes, while in MSMC-IM these two effects are separated (S3 Fig).

Deep ancestry in Africa

We applied our model to 30 high coverage genomes from 15 world-wide populations from the SGDP dataset [12] (S1 Table) to analyze global divergence processes in the human past (Figs 4–6). When analyzing the resulting pairwise migration rate profiles, we find that several population pairs involving African populations exhibit by far the oldest population structure observed in all pairwise analyses. We find that in all population pairs involving either San or Mbuti, the main separation process from other populations dates to between 60-400kya, depending on the exact pair of populations (see below), but with small amounts reaching back to beyond a million years ago, as seen by the non-zero migration rates around that time (Fig 4A, S4 Fig), and the cumulative migration probability, $M(t)$, (Fig 4B) which has not fully reached 1 until beyond a million years ago. Following the interpretation of $M(t)$ as discussed above with the archaic-admixture simulation scenario, we can infer that in pairs involving San or Mbuti, at least around 1% of ancestry can be attributed to lineages of ancestry that have diverged from the main human lineage beyond 1 million years ago (see also Fig 7, discussed further below). The genetic separation profile in pairs involving Mbuti and San is, beyond the extraordinary time depth, not compatible with clean population splits (as seen in simulations, Fig 2A) or simple scenarios of archaic admixture, but instead shows evidence for multiple or ongoing periods of gene flow between (unsampled) populations. Between Mbuti and other African populations except San, we find three distinct phases of gene flow. The first peaks around 15kya, compatible with relatively recent admixture between Mbuti and other African populations. The second phase spans from 60 to 300kya, reflecting the main genetic separation process, which itself looks complex and exhibits two peaks around 80-200kya thousand years ago. The third and final phase, including a few percent of lineages from around 600kya to 2 million years ago, likely reflects admixture between populations that diverged from each other at least 600kya. In pairs that include San, the onset of gene flow with other populations is more ancient than with Mbuti, beginning at around 40kya and spanning until around 400kya in the main phase, and then exhibiting a similarly deep phase as seen in Mbuti between 600kya and 2 million years ago. We confirm that this deep divergence is robust to phasing strategy (see below) and filtering (see Methods). We also replicated this signal using an independent dataset [14] (S5 Fig). An exception to these signals seen with San and Mbuti are pairs involving Karitiana, which do not exhibit such deep divergence. This is likely due to the strong genetic drift present in Karitiana, and the low heterozygosity in that population [12], which may shadow deep signals.

Apart from the deep structure seen with Mbuti and San, we find the second-most deep divergences between the West African Yoruba, Mandenka and Mende on the one hand, and French on the other (Fig 5A, S4 Fig, Fig 7 discussed further below), based on the time when M (t) reaches 99%. This might be consistent with recent findings of archaic ancestry in West-Africans [15,16], although it is not clear why the signal is primarily seen with French, and less consistently with Asian populations (Yoruba/Han have deep divergences, as well as Mende/Dai and Mandenka/Dai, but not other West-African/Asian combinations). Finally, pairwise analyses among Mende, Mandenka and Yoruba (Fig 4A, S4C, S4E and S4F Fig) exhibit a very

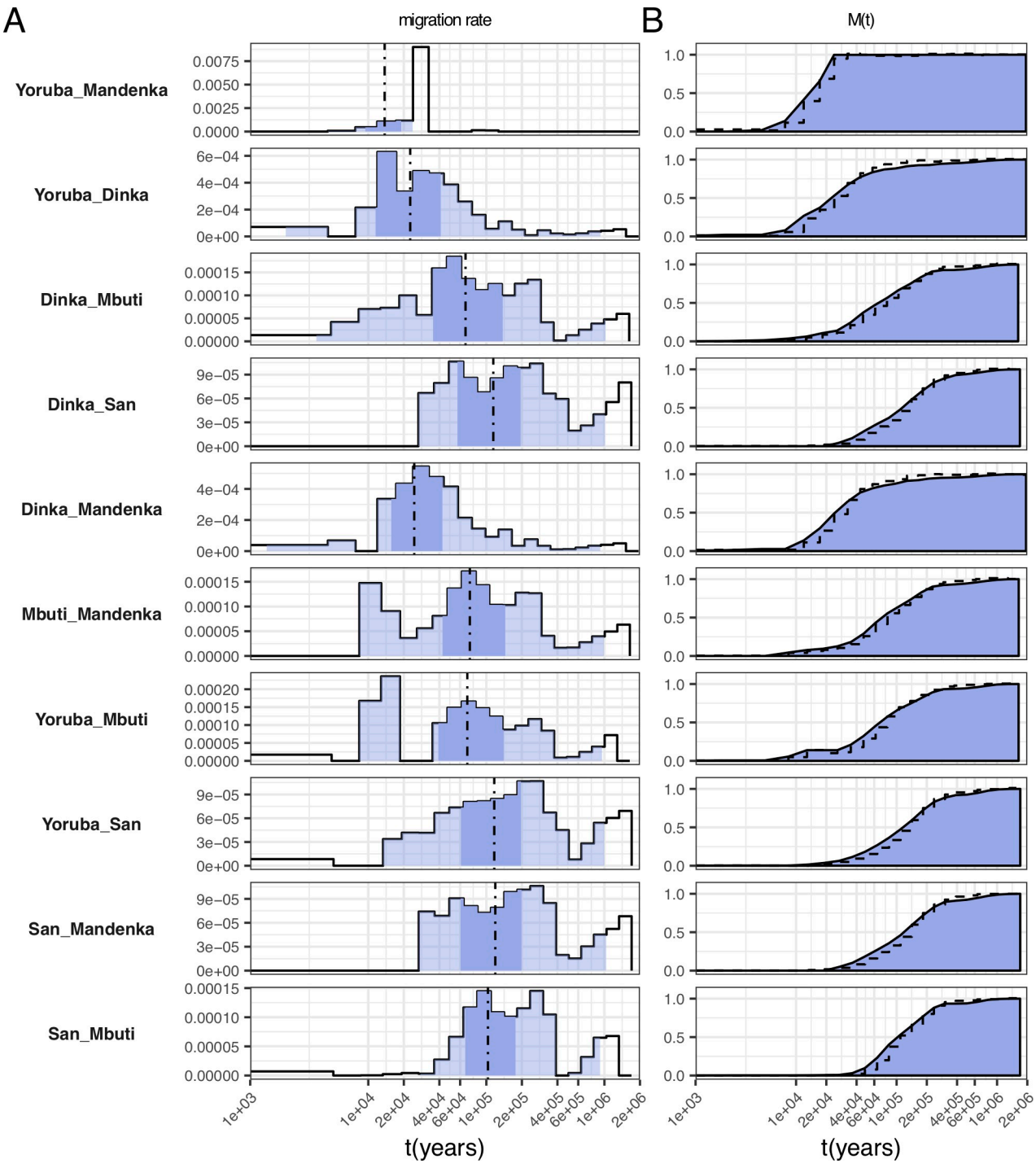


Fig 4. Migration rate profiles for selected pairs of African populations. (A) Migration rates. Dashed lines indicate the time point where 50% of ancestry has merged, and shading indicates the 1%, 25%, 75% and 99% percentiles of the cumulative migration probability (see Fig 2). (B) Cumulative migration probabilities *M(t)*. Dashed lines indicate the relative cross coalescence rate obtained from MSMC2, for comparison. See S4 Fig for the full set of figures.

https://doi.org/10.1371/journal.pgen.1008552.g004

recent migration profile, which appears to span up to about 20kya but not older, which is at odds with a recent finding of basal African ancestry present to different degrees in Mende and

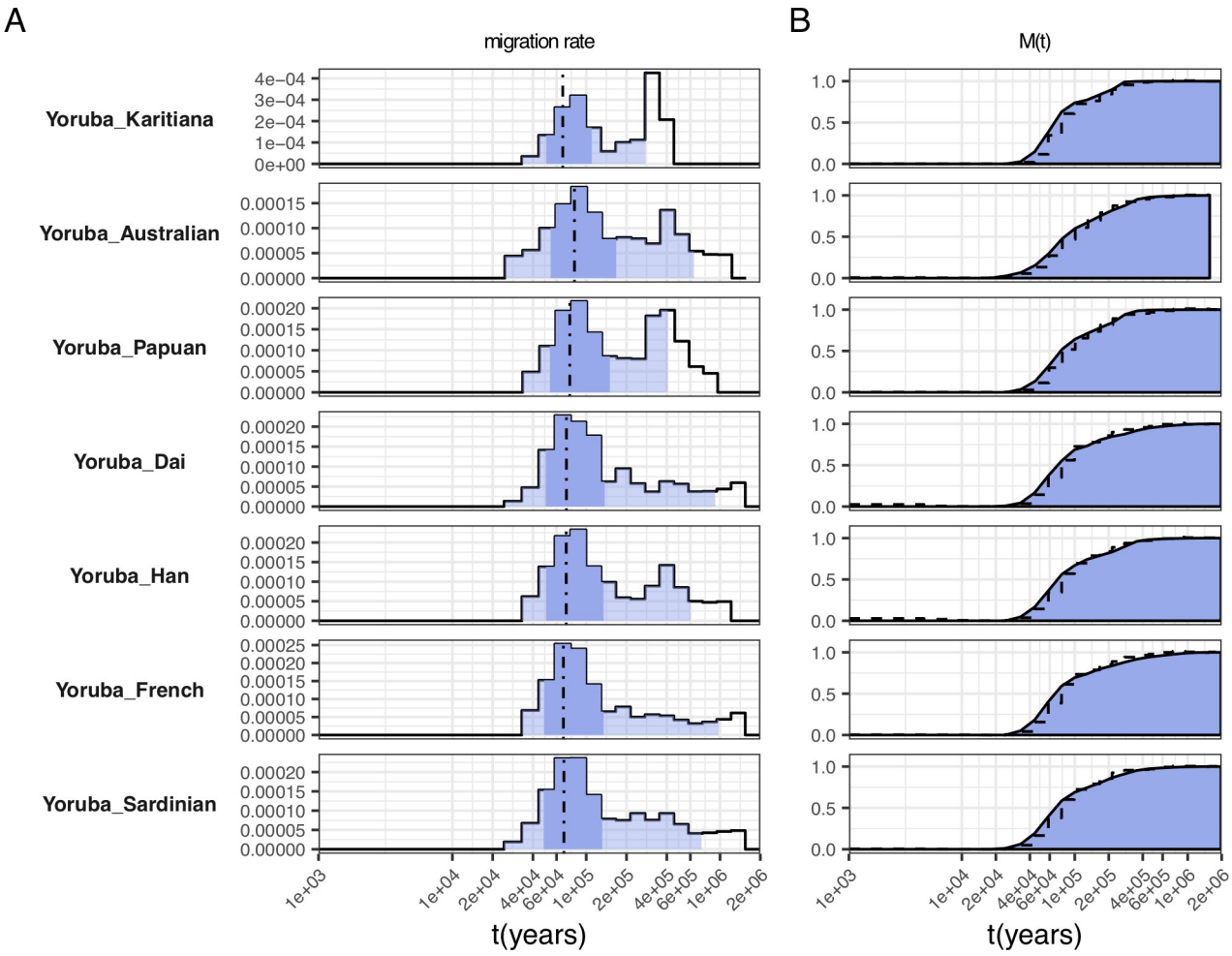


Fig 5. Selected migration profiles between Yoruba and 7 non-African populations. (A) Migration rates. Dashed lines indicate the time point where 50% of ancestry has merged, and shading indicates the 1%, 25%, 75% and 99% percentiles of the cumulative migration probability (see Fig 2). (B) Cumulative migration probabilities *M(t)*. Dashed lines indicate the relative cross coalescence rate obtained from MSMC2. See S4 Fig for the full set of figures.

https://doi.org/10.1371/journal.pgen.1008552.g005

Yoruba [17]. However, that signal may be too weak to be detected in our method, which is based on only two individuals per population.

Complex divergence between African and Non-African populations

Compared to the separation profiles between San or Mbuti and other populations, separations between other Africans and non-Africans look relatively similar to each other, with a main separation phase between 40 and 150kya, and a separate peak between 400 and 600kya (Fig 5 and S4 Fig). The first, more recent, phase plausibly reflects the main separation of Non-African lineages from African lineages predating the "out-of-Africa" migration event, and coinciding with the major population size bottleneck observed here (S6 Fig) and previously [3,4] around that time period. Signals more recent than about 60kya likely reflect the typical noisy spread of MSMC-estimated coalescence rate changes observed previously [4]. The second peak of migration, between 400 and 600kya likely reflects Neandertal and/or Denisovan introgression into non-Africans. The age of that peak appears slightly more recent than, although

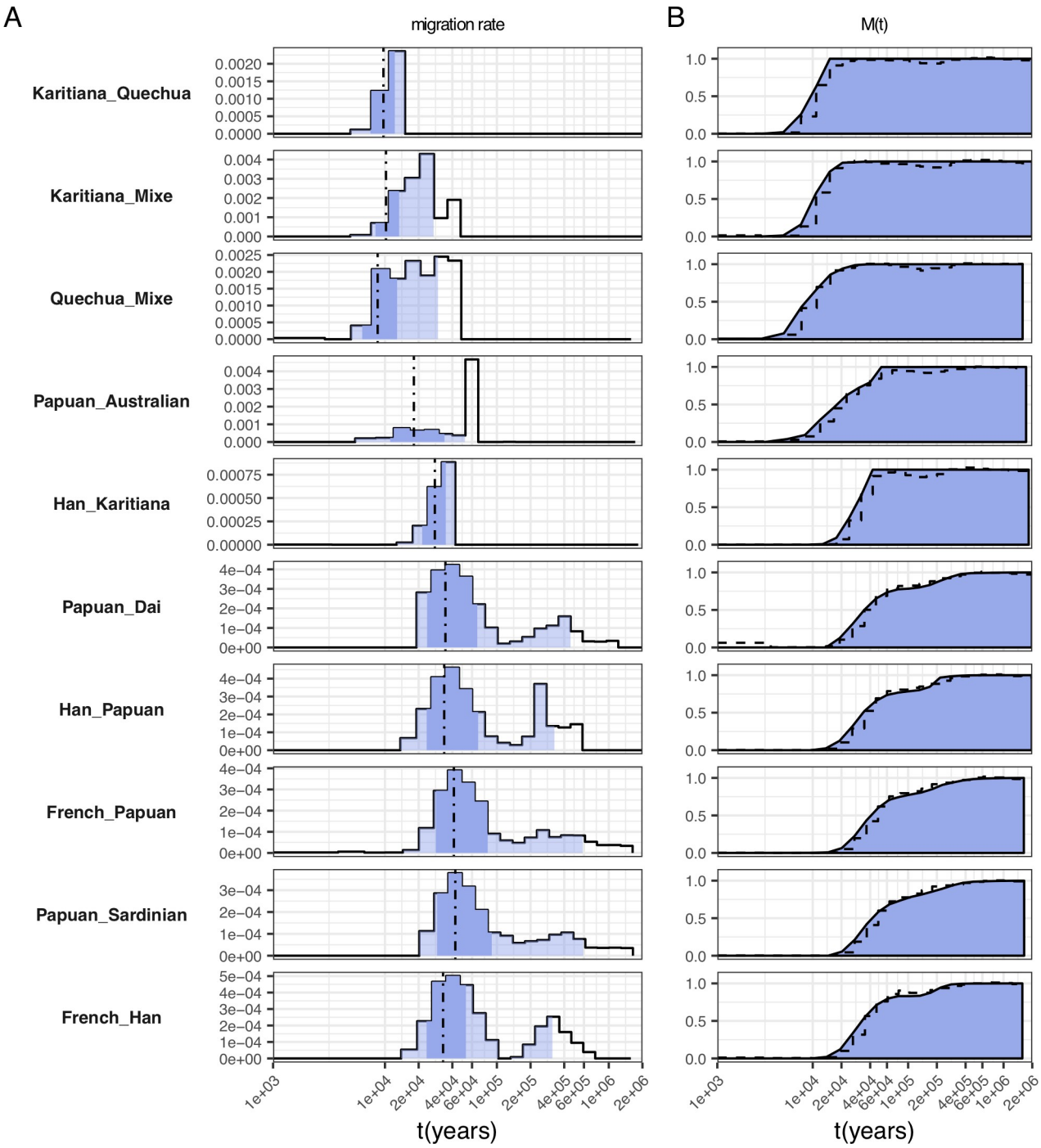


Fig 6. Selected migration profiles within non-African populations. (A) Migration rates. Dashed lines indicate the time point where 50% of ancestry has merged, and shading indicates the 1%, 25%, 75% and 99% percentiles of the cumulative migration probability (see panel B). (B) Cumulative migration probabilities M(t). Dashed lines indicate the relative cross coalescence rate obtained from MSMC2. See S4 Fig for the full set of figures.

https://doi.org/10.1371/journal.pgen.1008552.g006

overlapping with, previous split time estimates of those two Archaic groups from the main human lineage at 550-765kya [14]. However, our simulation with archaic admixture with bottleneck (Fig 2D), shows that our model tends to underestimate the archaic split time in the

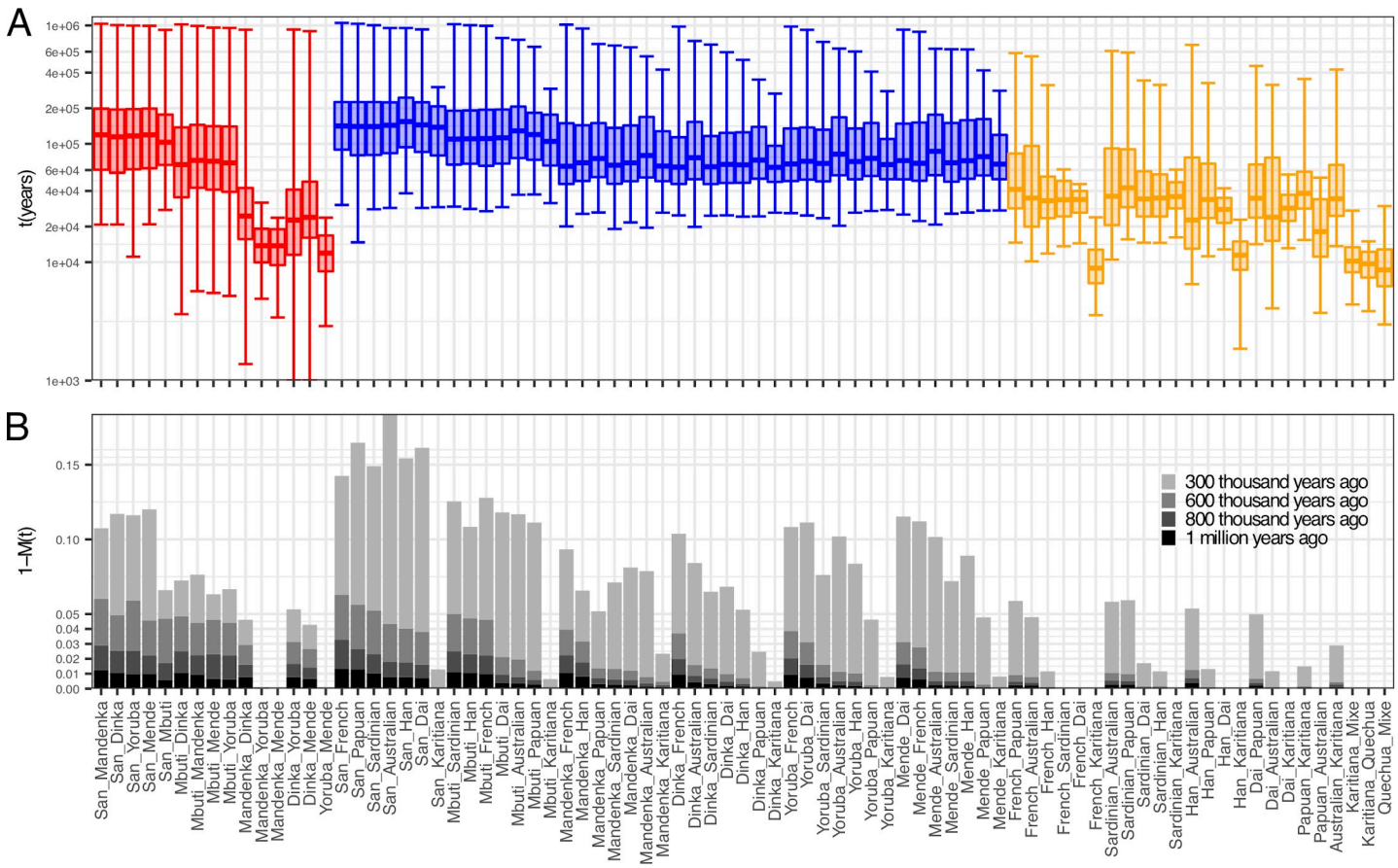


Fig 7. Summary profiles for divergence processes for 81 pairs of populations from 15 populations. (A) Boxes show the 25% to 75% quantiles of the cumulative migration probability *M(t)*, with bi-directional elongated error bars representing 1% and 99% percentiles. Colorcode: Red for African/African, blue for African/Non-African and orange for Non-African/Non-African pairs. (B) Barchart showing the amount of ancestry due to lineages older than 300, 600, 800kya and 1 million years ago, based on the cumulative migration probability *M(t)*.

https://doi.org/10.1371/journal.pgen.1008552.g007

presence of population bottlenecks as is the case for non-African populations [18–20]. In favor of the hypothesis that this second peak is caused by archaic lineages that have contributed to non-Africans is the fact that in all pairs of Papuans/Australians vs. Yoruba/Mende/Mandenka or Dinka, the second peak is particularly pronounced. This fits the archaic contribution hypothesis, since Papuans and Australians are known to have among all extant human populations the highest total amount of ancestry related to Neanderthals and Denisovans.

We investigated previous observations of potential ancestry from an earlier dispersal out of Africa, present in Papuan and Australian genomes [12,13,21]. Previously, one line of evidence for such a signal was based on shifts of relative cross coalescence rate curves between some Africans and Papuans or Australians on the one hand compared to curves with Europeans or East Asians on the other. With MSMC-IM we can compare these curves more quantitatively. While we were able to replicate this slight shift of relative cross coalescence rate or *M(t)* midpoint-based split times from African/Eurasian pairs to African/Australasian pairs reported in Ref. [21] using MSMC and Ref. [13] using MSMC2, we find that the estimated migration profiles of these pairs are very similar (S7 Fig), with a main separation midpoint around 70kya and a second older signal beyond 200kya, consistent with both Australasians and other Non-Africans being derived from a single genetic ancestral population without a more basal contribution to Australasians [12,13]. We conclude that the shift in the relative cross coalescence

rate curve appears to be consistent with being caused by the higher amount of archaic ancestry present in Papuans and Australians. We note, however, that different separation events are not distinguishable in MSMC-IM when they are temporally close to each other, as we saw in the *split-with-migration-scenario* (Fig 2B).

Separations outside of Africa

All separations outside of Africa are younger than separations between Africans and Non-Africans, as expected (Fig 6, S4 Fig). The deepest splits outside of Africa are seen in pairs of Papuans or Australians with other Eurasians, in which the first peak of migration is seen at 34kya, corresponding to the early separation of these populations' ancestors from other non-African populations after the out of Africa dispersal. In these pairs we see a second peak around 300kya, likely corresponding to the known Denisovan admixture in Papuans and Australians [13,22]. This is too recent for divergence time estimates between Denisovans and modern humans [14], which again is consistent with the underestimate seen in simulations with bottlenecks. Surprisingly, we see a similar second peak between French and Han, which is consistent with cross-coalescence rate features in previous observations [4,12] but of unclear cause. Consistent with the hypothesis that the second peak seen in Australasian/Eurasian pairs corresponds to Denisovan admixture, we do not see a second peak in the migration profile between Papuans and Australians, confirming that the gene flow likely occurred into the common ancestor of Australians and Papuans [13]. The migration profile between Papuans and Australians shows a main separation between 15-35kya.

The second deepest splits in Non-African populations are seen between East Asian and European populations, which occur mostly between 20 and 60kya (cumulative migration probability midpoint at 34kya), followed by separations between Asian and Native American populations, between 20 and 40kya (midpoint at 28kya). The latter likely also reflects Ancestral North Eurasian ancestry in Native Americans [23], which is more closely related to Europeans than to East Asians, thereby pushing back the separation seen between East Asians and Native Americans. Finally, the most recent splits are seen between populations from the same continent: Dai/Han split around 9-15kya (midpoint 11kya), French/Sardinian around 7-13kya (midpoint 9kya) and within Native Americans around 7-13kya (midpoint 10kya) (Fig 6, S4 Fig).

To visualize the depth of ancestry in each population pair, we summarized all pairwise analyses by percentiles of the cumulative migration probability $M(t)$ (Fig 7). Largely, Non-African pairs (orange) have their main separation phase, with the cumulative migration probability between 25% and 75%, between 20 and 60kya, with some more recently diverged pairs within continents. In contrast, African pairs (red) have their main phase largely between 60 and 200kya, with some notable exceptions of more recently diverged populations, and with the notable tail (99% percentile) up to 1 million years and older. Between Africans and Non-Africans, divergence main phases are largely within a similar window of 60-200kya as in African pairs, with three notable groups: divergence of Non-Africans from San falls between 80-250kya, from Mbuti between 70-200kya, and from other Africans between 50-150kya. To highlight the amount of ancestry contributed asymmetrically to one of the two populations from unsampled populations that diverged from the human lineage in the deep past (so-called archaic lineages), we show the distance of the cumulative migration probability from 1, 1-$M(t)$, at different deep time points (Fig 7B). As described above, the deepest signals are seen in pairs involving San or Mbuti, reaching 3% of ancestry contributed from lineages that diverged at least 800kya, and around 1% of ancestry from lineages that diverged at least 1 million years ago. Similarly deep levels are seen in specific pairs involving French, in combination with the

West African Mende, Mandenka and Yoruba and the East African Dinka, and for pairs Mende/Dai and Mandenka/Han, as discussed above.

Robustness to phasing and processing artifacts

MSMC2 (like MSMC) requires phased genomes for cross coalescence rate estimation, and we therefore rely on statistical phasing within the SGDP dataset, for which different strategies are possible. To compare the effect of selecting such phasing strategy, we generated phased datasets using eight different phasing strategies with three phasing algorithms (SHAPEIT [24], BEAGLE [25], EAGLE [26]). We included genotype calls from 12 individuals with previously published physically phased genomes [12] and then used those genomes to estimate the haplotype switch error rate. Among eight phasing strategies, SHAPEIT2 [24], without the use of a reference panel, but including information from phase-informative reads [27], resulted in the lowest switch error rate per kb (and per heterozygous site; S8 Fig). Overall, switch error rates are higher in African populations, likely due to lower linkage disequilibrium, higher heterozygosity and relatively limited representation in the SGDP. To test how sensitive MSMC-IM is to different phasing strategies, we tested four phasing strategies on four different pairs of populations with evidence for extremely deep ancestry (Methods). We find that the migration profile from MSMC-IM is very similar for different phasing strategies. In particular, we find that the very deep signal seen in population pairs involving San and Mbuti is reproduced with different phasing strategies with and without a reference panel (S9A Fig). In a similar way, we confirmed the robustness of that signal with respect to choosing different filter levels (S9B Fig) and with respect to removing CpG sites, which are known to have elevated mutation rates (S9C Fig). We also explored to what extent switch errors affect our estimates using simulated data (S10 Fig), and confirmed robustness with respect to variation in recombination rates, which are assumed to be constant along the genome within MSMC2 but vary in reality (S11 Fig). Finally, to test internal consistency, we tested how well MSMC-IM was able to infer back its own model. We used the estimated migration rates and population sizes from eight population pairs (see Methods), and simulated genomic data under their inferred models. As shown in S12 Fig, the estimated migration patterns from the simulated and the real data are indeed very similar, including the deep signals seen in pairs with San and Mbuti.

Given the superiority of the read-aware phasing strategy with SHAPEIT without a reference panel [27,28] (S8 Fig), we used this method in all of our main analyses. However, even with this phasing strategy, the switch error rate is high in populations that are not well represented in the dataset. In case of indigenous Australians, the phasing quality is among the worst in the dataset (S8 Fig), arguably because the SGDP dataset contains only two Australian individuals (compared for example to 15 Papuans). To improve phasing in Australians specifically, we generated new high coverage genomic data for one of the two Australians in the SGDP dataset using a new library with longer read-pair insert sizes (see Methods). Using these additional reads reduced the switch error rate from 0.038/kb to 0.032/kb. (S8 Fig, blue isolated dot for Australian3). We ran MSMC2 on the long-insert Australian data, as well as the standard phased data, combined with one diploid genome from each of the other world-wide populations analyzed in this study. The inferred migration profiles from MSMC-IM (S13 Fig) for Non-African population pairs involving the long-insert phased Australian genome do not seem to be affected by the phasing method (S13 Fig). The migration profile from pairs of Africans versus the long-insert phased Australian tend to be slightly younger, but also show deeper structure in Dinka/Australian, compared to the same pair using the *shapeit_pir* phasing method, which uses phase-informative sequencing reads to improve phasing accuracy (Methods). Note that these migration rate densities exhibit more noise than the ones used in our

main analysis (S4L Fig), since they are based on only one individual per population, while the main analyses are based on two individuals per population. The main separation between Papuan and Australian remains at 15-35kya, as shown in the migration profile from both phasing strategies, very close to the estimates from 8 haplotypes in the main analysis (S4L Fig), and earlier than the previous estimates of 25-40kya [13].

Similar to the procedure introduced for PSMC [3], we use a block-bootstrap approach to assess statistical uncertainty of our method. We find that there is very little uncertainty around MSMC-IM's migration rate estimates (S14 Fig) based on these bootstrap-estimates. This should be taken with caution, though, since the bootstrap is only able to address the uncertainty caused by randomness in the data, not by systematic biases. We know that MSMC typically "smears out" sudden changes in coalescence rates, which is due to the wide variance in local estimates of coalescence times, and this type of error is not revealed by the bootstrap. It does, however, give high confidence to specific results, such as estimates of archaic ancestry between 1 and 20% as seen in Fig 3C. According to our bootstrap test (S14 Fig), the cumulative migration probability M(t) does hardly vary at all in bootstrap replicates, so estimates of deep ancestry fractions such as the ones shown in Fig 3C and Fig 6B for real data, are very accurate.

Discussion

We have presented both a novel method MSMC-IM for investigating complex separation histories between populations, and an application of that method to human genomes, revealing new insights into the complex separations and deep ancestry in African populations. MSMC-IM extends MSMC2 by fitting an IM model to the estimated coalescence rates, which allows us to characterize the process of population separation via a continuous migration rate through time. In contrast to the established approach of using the relative cross coalescence rate directly from MSMC2, our new approach interprets coalescence rates more quantitatively. In a recent study a similar approach has been used to fit an IM model to PSMC estimates to estimate population split times and post-split migration rates in a more strictly parameterized model [29]. We found here that a continuous IM model without an explicit split time better fits the estimated coalescence rates from MSMC2, which are continuous themselves and thus lead to a more gradual concept of population separation. This absence of an explicit population split time distinguishes our approach from many previous models [5,8,9] and allows us to detect new signals of temporal population structure without specifying population phylogenies or admixture graphs from prior knowledge or via inference.

A showcase example for such new insights are the traces of extremely deep population structure seen in our analysis of African population pairs. The fact that San and Mbuti exhibit the deepest branches in the human population tree is itself not surprising given previous analyses [30–34], but the extraordinary time depth displayed in this analysis has to our knowledge not been reported before. This deep structure—albeit only making up 1% of ancestry—is far older than the oldest attested fossil records of anatomically modern humans, considering the East-African fossils of Omo Kibish and Herto 160-180kya [34–36] and the skull from Jebel Irhoud recently re-dated to around 300kya [37]. Any admixture from an archaic population that diverged from the main human lineage more than 600kya would produce such a signal. This is the case, for example, for the so-called "super-archaic" population that was inferred to have admixed into Denisovans [14] and was estimated to have diverged from the lineage leading to modern Humans, Neanderthals and Denisovans between 1.1 and 4 million years ago. Given this finding outside of Africa, it is perhaps not surprising that such deep archaic population structure existed also in Africa.

However, our signal of archaic population structure in Africa reveals more complexity than expected under the standard model of archaic introgression, in which two divergent populations admix with each other, creating a distinct pattern of deep ancestry in the genomes of the target population. Detecting such patterns in the genome would require a sufficient sequence divergence between non-introgressed and introgressed genomic segments and sufficiently long introgressed segments (as detected by the S^* statistic or extensions of it [15,16]). This is the case if the majority of ancestry between the two intermixing species has been isolated for hundreds of thousands of years, with a relatively recent introgression time (comparable to the time of the Neanderthal introgression). Such a scenario would then be seen as a bimodal pattern in the migration profile, as shown in our simulations (Fig 2C). What we see, however, in the migration profiles between San and Mbuti with other African populations, is not a bimodal pattern, but a more continuous distribution. This would emerge under a model of repeated isolation and partial admixture of two or more archaic species or populations that exist in parallel for a long time. Under such a scenario, genomes are not a two-way mixture between introgressed and non-introgressed regions, but a mosaic of ancestry lines merging at a range of different split times. Since much of the introgression would then be attributed to very ancient events, these segments would be too short for methods such as S^* to be detected as archaic ancestry, which may be the reason why the deep signals reported here have not been reported before for San and Mbuti, in contrast to Non-Africans and West Africans [15,16].

While the continuous model in MSMC-IM adds significantly to previous approaches to estimating population separations, one drawback is that it is currently limited to only two populations at a time. While this limit is partially technical—MSMC2 cannot be scaled to arbitrary numbers of genomes—the more severe problem is a conceptual one. It is not obvious how to use the concept of continuous-time migration rates and non-sharp population separations to more than two populations. One possibility are graph models, as they are used in admixture graphs [38], but it is unclear how to make such models fully continuous, as is our current migration rate and cumulative migration probability for two populations. An important direction for future work is to achieve a generalization of the continuous concept of population separation to multiple populations, which might help to better understand and quantify the processes that shaped human population diversity in the deep history of our species.

Materials and methods

MSMC2

MSMC-IM is based on MSMC2 (first described and used in Ref. [13]) as a method to estimate pairwise coalescence rates from multiple genome sequences. The MSMC2 method is summarized in a self-contained way in S1 Text. MSMC2 is similar to MSMC [4], but instead of analyzing multiple genomes simultaneously modelling the first coalescence event, it uses the pairwise model in sequence on all pairs of haplotypes to obtain a composite likelihood of the data given a demographic model. The demographic model itself (consisting of a piecewise constant coalescence rate) is then optimized via an Expectation-Maximization algorithm similarly to MSMC and PSMC [3]. For cross-population analyses, we use MSMC2 to obtain three independent coalescence rate estimates: two coalescence rates through time within each population, named $\lambda_{11}(t)$ and $\lambda_{22}(t)$, respectively, and one coalescence rate function for lineage pairs across the population boundary, named $\lambda_{12}(t)$ (S1 Text).

MSMC-IM model

MSMC-IM then fits a two-island model with time-dependent population sizes $N_1(t)$ and $N_2(t)$, and a time-dependent continuous symmetric migration rate $m(t)$ to the estimated coalescence

rates, which essentially is a re-parameterization from the triple of functions $\{\lambda_{11}(t), \lambda_{12}(t), \lambda_{22}(t)\}$ to a new triple of functions $\{N_1(t), N_2(t), m(t)\}$ (S1 Text). To fit the island-model to the coalescence rates, we first use the coalescence rates to compute a probability density for times to the most recent common ancestor (tMRCA), as illustrated here for rate $\lambda_{11}(t)$:

$$P^{MSMC}(t|s_0 = S_{11}) = \lambda_{11}(t)e^{-\int_0^t \lambda_{11}(t')dt'}$$

Here, S_{11} denotes the starting state where both lineages are present in population 1. We then use an approach by Hobolth et al 2011 [39] to compute this density for the three starting states $s_0 = \{S_{11}, S_{12}, S_{22}\}$ under an IM model, denoted $P^{IM}(t|s_0)$, using exponentiation of the rate matrix of the underlying IM-Markov process that governs the state of uncoalesced and coalesced lineages in two populations connected by a time-dependent migration rate (see S1 Text). The fitting process of the IM model to the probability density computed from MSMC2 is done by minimizing the Chi-square statistics:

$$\chi^2 = \sum_{i=1}^{n_T}\left[\sum_{s_0\in\{S_{11},S_{12},S_{22}\}}\frac{\left(P^{IM}(t_i|s_0) - P^{MSMC}(t_i|s_0)\right)^2}{P^{MSMC}(t_i|s_0)} + \beta_1\int_0^\infty m(t_i)\,dt + \beta_2\left(\frac{N_1(t_i) - N_2(t_i)}{N_1(t_i) + N_2(t_i)}\right)^2\right]$$

where n_T denotes the number of time segments, and the t_i denote the boundaries of the discrete time segments. The second and third term in the formula are regularization terms to avoid overfitting, with β_1 restricting migration rates and β_2 pushing the two population sizes $N_1(t)$ and $N_2(t)$ close to each other. The strength of this regularization can be controlled via a user-defined parameter in our program. We sum over Chi-square statistics over n_T time intervals with i representing the time index in the formula. For the three simulation scenarios and all pairs of real data, we used a regularization value of $\beta_1 = 10^{-8}$, $\beta_2 = 10^{-6}$. Regularization is necessary because the reparameterization introduced by MSMC-IM overspecifies the model at times when the two populations are fully merged. For that same reason, we plot estimated migration rates in all figures only up to a value of $M(t) = 0.999$, since migration rate estimates beyond that point are essentially arbitrary, as lineages have already been fully randomized between the two populations. We also restrict the estimated population sizes to 10,000,000 in practice.

We implemented the MSMC-IM model as a python command line utility that takes the MSMC or MSMC2 output files as input. The program is available at: https://github.com/wangke16/MSMC-IM.

Simulations

We used *msprime* [40] for all simulations in this paper. In the three series of simulation scenarios mentioned above, we simulated four diploid genomes composed of 22 chromosomes each of length 100Mbp from two populations, assuming a constant population size 20,000 for every population. The recombination rate we used here is 10^{-8} per generation per bp, and the mutation rate is 1.25×10^{-8}.

In the zig-zag simulation (S1 Fig), we simulated a series of exponential population growths and declines for two, four and eight haplotypes, each changing between 3,000 and 30,000 in exponentially increasing time intervals, with the same simulation parameters as specified in Ref. [4] and Ref. [3] to ensure comparability with these previous publications. In particular, this simulation involved a lower recombination rate (0.3×10^{-8}) than the main simulations, justified in Ref. [4] as the inferred recombination rate from real data using PSMC'. The reason for it being lower than the true recombination rate (close to 10^{-8}, as used in the main simulations

above), is that MSMC (and MSMC2) infers an "effective recombination rate", which is a non-trivial average over the variable recombination landscape across the human genome.

We also conducted a number of simulations based on MSMC-IM inference from real data (S12 Fig). We took the estimates on migration rates and population sizes from MSMC-IM (S2 Table) for eight pairs of worldwide populations (San/Mbuti, San/Dinka, San/French, Mbuti/French, Yoruba/French, Yoruba/Papuan, French/Han, Papuan/Australian), as the input parameters in our simulation, and simulated 2.2Gb genomes on 8 haplotypes for each case. The recombination rate we used here is 10^{-8} per generation per bp, and the mutation rate is 1.25×10^{-8}.

Processing genomic Data

For the results shown in Figs 4–7, we used 30 high coverage genomes from 15 cross-continental modern populations in the SGDP dataset [12], with two diploid genomes from each population for running MSMC2 and MSMC-IM (S1 Table). Only the autosomal genome was used for this analysis. We ran pairwise analyses for 13 populations (excluding Quechua and Mixe) and pairwise comparisons within three native American populations (81 population pairs in total). We downloaded the *cteam-lite* dataset of from the website: http://reichdata.hms.harvard.edu/pub/datasets/sgdp/, in the *hetfa*-format where all sites are represented by an IUPAC encoding representing diploid genotypes, along with individual masks recording the quality of the genotype calls. We converted the *hetfa* mask files (.ccompmask.fa.rz) to zipped bed format though two steps: first we uncompressed the *hetfa* mask files using "*htsbox razip -d -c*" (https://github.com/lh3/htsbox), and then converted the uncompressed mask files (.ccompmask.fa) to zipped bed format by an in-house python script adapted from the *makeMappabilityMask.py* script in *msmc-tools* (www.github.com/stschiff/msmc-tools). The *cteam-lite* masks encode quality using an integer-range from 0 to 9 (reflecting increasing stringency) and "N" to represent missing data. For our analysis, we included all sites that were non-missing, i.e. have a minimum quality level of 0.

Following the processing introduced in PSMC [3] and MSMC/MSMC2 [4], beyond the individual masks we also use a universal mask to reflect overall mappability and SNP calling properties along the human genome. We used the universal masks defined in Supplementary Info 4 from Ref. [12] (and available for download at https://github.com/wangke16/MSMC-IM/tree/master/masks) as additional negative masks denoting genomic regions to be filtered out.

Beside the genome-wide mask files for each individual, we obtained variant data as made available on the SGDP project website (https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data/). Due to the specifics of how that dataset was generated, only segregating sites at positions where the Chimpanzee reference genome has non-missing data are included. To balance this missingness based on the Chimpanzee reference genome for MSMC, we included an additional mask in our preprocessing, which reflected non-missing regions in the Chimpanzee reference sequence. For others to reproduce our analysis, we provide this chimp mask on the MSMC-IM *github* repository (https://github.com/wangke16/MSMC-IM).

We phased the data using SHAPEIT2 (v837) [24], Beagle4.0 (r1399) [25] and EAGLE2 (version 2.3) [26]. We first phased the data using each algorithm both with and without a reference panel (here we used the 1000 Genomes Phase 3 reference panel as recommended in the Shapeit2 documentation). When using a reference panel, all three methods are only able to phase sites that are represented in the reference panel. Therefore, we removed sites not in the reference panel, phased, adding the removed sites back as unphased, and then ran a second round of phasing using Beagle4.0 and the "usephase = true" option, which allows us to phase the unphased sites in data that is already partially phased. Finally, we also phased using SHAPEIT2

without a reference panel, but using the read-aware phasing strategy [27]. This uses the fact that two SNPs found on the same (paired) read must be in phase. The switch error of each of these phasing strategies, evaluated by comparison with the experimentally phased data generated for the same samples [12] is shown in S8 Fig.

Finally, we generated a long-insert library from one of the two Australian DNA samples analyzed in SGDP [12], with a median insert size of 3.3kbp. These data are available at the European Nucleotide Archive under accession number ERX1790596 (https://www.ebi.ac.uk/ena/data/view/ERX1790596). We used this data to improve the phasing quality for this Australian individual. As shown in S8 Fig, this strategy indeed reduced the switch error rate for this Australian individual from 0.036/kb to 0.032/kb.

Running MSMC-IM

Unlike MSMC, which reports these three rates in a single analysis step, in MSMC2 we run the three estimations for $\lambda_{11}(t)$, $\lambda_{12}(t)$ and $\lambda_{22}(t)$ independently from each other, using a different selection of haplotype pairs in each case. We base most of our analyses on 4 diploid individuals (unless indicated otherwise), for which we prepared joint input files for each chromosome, consisting of 8 haplotypes each. We then chose the pairs to be analyzed using the "-I" option in MSMC2. For coalescence rate $\lambda_{11}(t)$, we used "-I 0,1,2,3", which instructs MSMC2 to iterate through all six possible haplotype pairs among the four haplotypes from the first population. Likewise, to estimate $\lambda_{22}(t)$, we used "-I 4,5,6,7". Finally, to obtain estimates of the coalescence rates across populations, $\lambda_{12}(t)$, we used "-I 0–4,0–5,0–6,0–7,1–4,1–5,1–6,1–7,2–4,2–5,2–6,2–7,3–4,3–5,3–6,3–7", iterating through all sixteen possible haplotype pairings between the four haplotypes in each population. MSMC-IM requires a single input file containing all three coalescence rate estimates, similar to the output generated by the original MSMC program. A script *combineCrossCoal.py* is provided on the *msmc-tools github* repository (http://www.github.com/stschiff/msmc-tools), to generate the combined output file from the three output files of the three MSMC2 runs for a pair of populations.

With the combined MSMC2 output as input, we run MSMC-IM model by "***MSMC_IM.py pair.combined.msmc2.txt***". Also, the time pattern needs to be specified, which is by default ***1*2+25*1+1*2+1*3*** as the default in MSMC2. In the output, MSMC-IM will rescale the scaled time in MSMC2 output by mutation rate 1.25e-8 into real time in generations, and report symmetric migration rates and $M(t)$ in each time segment.

Robustness tests

Phasing Strategy: We tested the robustness of our findings by applying four different phasing strategies—*beagle*, *shapeit*, *shapeit_ref_all* to *shapeit_pir* to four pairs of populations in the SGDP dataset (San/Mbuti, San/Yoruba, San/French, Mbuti/French). Here, *beagle* and *shapeit* denote phasing with no reference panel, *shapeit_ref_all* denotes phasing with a reference panel (1000 Genomes phase 3, with sites not in the reference panel phased with Beagle) and *shapeit_pir* denotes no reference panel but including phase-informative reads (S9 Fig).

Filtering: We explored the impact of mask filtering levels using San/French and Mbuti/French in the SGDP dataset, by varying the stringency of the filtering between levels 0, 1, 3, 5 (S9 Fig).

CpG islands: We conducted San/French and Mbuti/French runs with removed CpG sites. For this, we generated a mask including all positions of Cytosines and Guanines in CpG dinucleotides, Thymines in TpG dinucleotides, and Adenosines in CpA dinucleotides in the human reference genome hg19, and used those positions as negative mask when preparing the

MSMC input files. This mask can be found in the *github* repository (https://github.com/wangke16/MSMC-IM).

Simulated switch errors: To explore the impact of switch errors, we added artificial switch errors at rates ranging from 5e-6 to 5e-4 per base pair in four different simulation scenarios—the *clean-split* scenario at 75kya, the *split-with-migration* scenario at 75kya, the *split-with-archaic-admixture* scenario at proportion 5%, the *split-with-archaic-admixture-bottleneck* at proportion 5%. As shown in S10 Fig, we found that the impact of switch errors on MSMC-IM's estimates is negligible up until switch errors of rate 5e-5.

Simulations with variable Recombination rates: In the four simulation scenarios selected above we simulated variable recombination rates using a human genetic map with variable recombination rates along the genome downloaded from (ftp://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2011-01_phaseII_B37/genetic_map_HapMapII_GRCh37.tar.gz). As shown in S11 Fig, MSMC-IM's estimates from using a real genetic map are consistent with estimates from using constant recombination rates.

Bootstrapping: We applied a block-bootstrap, similar to the approach described in ref. [3] to six pairs in the SGDP dataset (San/Mbuti, San/Dinka, Yoruba/French, French/Han, Yoruba/Papuan, Papuan/Australian) with 20 replicates for each (S12 Fig).

Independent Dataset: We tested our approach on 12 populations (24 genomes) from another dataset [14], which consists of different genomes available from http://cdna.eva.mpg.de/neandertal/altai/ModernHumans/. This dataset was processed independently using the pipeline in the *msmc-tools github* repository (http://www.github.com/stschiff/msmc-tools) i.e. SNPs and masks generated using *samtools* and *bamCaller.py*, with statistical phasing by *SHAPEIT2* with the 1000 Genomes reference panel, leaving sites not present in the reference panel as unphased. Results between the two datasets are very similar, with some differences observed in relation to highly drifted populations like Karitiana.

Supporting information

S1 Fig. MSMC and MSMC2 population size estimates from simulated data. To test population size inference capabilities of MSMC (A) and MSMC2 (B) applied to two, four and eight haplotypes, we simulated a series of exponential population growths and declines, each changing the population size by a factor ten. The true population size is shown as dark solid line. Compared to MSMC, MSMC2 recovers the population size well, and the resolution in recent times increases with the number of haplotypes. With two haplotypes, MSMC2 infers the population history from 10kya to 3 million years, whereas, with four haplotypes and eight haplotypes the resolution in recent times is extended to 3kya and 1kya years ago respectively.
(PDF)

S2 Fig. Cumulative migration probabilities from four simulation scenarios. This figure shows the same results as Fig 2, but showing M(t) instead of m(t). The scenarios are (A) the *Clean-split* scenario. (B) the *Split-with-migration* scenario, and (C) the *Split-with-archaic-admixture* scenario. (D) the *Split-with-archaic-admixture-and-bottleneck* scenario. For panel (C) and (D), we show results with alpha ranging from 0 to 1, instead of between 0 to 20% shown in Fig 2. The relative CCR is shown in step-wise dashed lines to be compared with M(t).
(PDF)

S3 Fig. Population size estimates from MSMC2 compared to MSMC-IM: We simulated $N_1(t)$ and $N_2(t)$ as constant 20,000 in top three different simulation scenarios, and simulated a severe bottleneck in $N_2(t)$ with a factor 30 between 40-60kya in the bottom

simulation scenario. The split time *T* is 75kya in all four cases, and all other parameters are the same as in Fig 2 and as indicated. As shown, the MSMC-IM estimates for $N_1(t)$ and $N_2(t)$ are close to the inverse coalescence rates, with relatively small effects caused by the migration rate in MSMC-IM which is absent from MSMC2.
(PDF)

S4 Fig. Pairwise migration profiles for 13 worldwide populations, involving San (A), Mbuti (B), Mandenka (C), Dinka (D), Yoruba (E), Mende (F), French (G), Sardinian (H), Han (I), Dai (J), Papuan (K), Australian (L), Karitiana (M). The relative CCR is shown in step-wise dashed lines to be compared with M(t). *See separate joint PDF file*.
(PDF)

S5 Fig. Migration profile of an independent dataset. Here we have analyzed 12 worldwide populations from Prüfer et al (2014) with independent data processing as described in Methods: San (A), Mbuti (B), Mandenka (C), Dinka (D), Yoruba (E), French (F), Sardinian (G), Han (H), Dai (I), Papuan (J), Australian (K), Karitiana (L). The relative CCR is shown in step-wise dashed lines to be compared with M(t). *See separate joint PDF file*.
(PDF)

S6 Fig. Estimated population sizes from MSMC2 for 15 worldwide populations. We show the estimates from MSMC using 8 haplotypes/4 individuals per population from the SGDP dataset.
(PDF)

S7 Fig. Testing for potential multiple out-of-Africa separations. Here we show analyses on the divergence of Papuans and Australians from Africans vs. other Non-African populations from Africans. We show the cumulative migration probability $M(t)$ in (A), and the migration rate $m(t)$ (B) for pairs of populations of Yoruba, Dinka and San with one non-African population as indicated.
(PDF)

S8 Fig. Switch error rates from eight phasing strategies. *beagle* and *beagle_ref_all* denote BEAGLE phasing without and with reference panel (here and below denoting the 1000 Genomes Phase 3 reference panel). *eagle* and *eagle_ref_all* represent EAGLE phasing without and with reference panel. *shapeit* and *shapeit_ref_all* represent SHAPEIT phasing without and with reference panel. *shapeit_pir* represents SHAPEIT phasing with phase-informative reads. *shapeit_pir_extra* represents SHAPEIT phasing with long-insert-size reads as additional phase informative reads, which was applied to B-Australian-3 only. See Methods for details.
(PDF)

S9 Fig. Impact of phasing and processing artifacts. We show (A) the impact of the phasing strategy using San/Mbuti, San/Yoruba, Mbuti/French and San/French as examples, (B) the impact of the filtering level for generating individual masks using San/French and Mbuti/French as example, and (C) the impact of removing CpG sites using San/French and Mbuti/French as example. See caption to S8 Fig for a description of the four phasing methods shown in (A).
(PDF)

S10 Fig. Impact of switch errors on simulated data. Here we selected the same four simulation scenarios used in S3 Fig, and added phasing switch errors ranging from 5e-6 to 5e-4 per base pair. The overall migration profiles remain relatively consistent for error rates between 5e-6 and 5e-5, with strong effects seen with rates higher than 5e-5, shifting the migration

profiles towards older times. (A) Clean split at 75kya. (B) Split at 75kya with symmetric migration between 10-15kya. (C) Split at 75kya with archaic admixture at 5%. (D) Split at 75kya with archaic admixture at 5% and bottleneck in one population.
(PDF)

S11 Fig. Impact of recombination rate on simulated data. Applying the same four simulation scenarios used in S3 Fig, we here used the genetic map estimated for the human genome (i.e. variable recombination rate across genome) instead of a constant recombination rate. Red lines represent our estimates from using a constant recombination rate 10^{-8} per generation per bp. (A) Clean split at 75kya. (B) Split at 75kya with symmetric migration between 10-15kya. (C) Split at 75kya with archaic admixture at 5%. (D) Split at 75kya with archaic admixture at 5% and bottleneck in one population.
(PDF)

S12 Fig. Migration profile on simulated pseudo-SGDP genomes. Green lines show the estimates we got from SGDP data for pairs shown on the left (as shown in S4 Fig), which are used as input parameters for the simulation. Red lines show the estimates from applying MSMC-IM on the simulated data. (A) Migration rates *m(t)*. (B) Cumulative migration probabilities *M(t)* and relative cross-coalescence rates.
(PDF)

S13 Fig. Impact of long-insert phasing on Australian population separation inferences. *M(t)* in quantiles is summarized here between a single Australian and a single individual from worldwide populations. Boxes show the 25% to 75% quantiles of *M(t)*, with bi-directional elongated error bars representing 1% and 99% percentiles. Red color represents the data phased using long-insert reads. Green color represents the standard phased dataset.
(PDF)

S14 Fig. Bootstrap tests. As shown in (A) migration rate m(t) and (B) Cumulative migration probability *M(t)*, the overall inferred profile for each pair is rather consistent across 20 replicates.
(PDF)

S1 Table. Analyzed samples and population labels from the SGDP dataset.
(XLSX)

S2 Table. MSMC2 results and MSMC-IM estimates for all pairs of SGDP populations analyzed in this paper, *see separate Excel file*. The columns reported are described within a legend included in the Excel file.
(XLSX)

S1 Text. Derivation of MSMC2 and MSMC-IM theory, *see separate PDF file*.
(PDF)

Author Contributions

Conceptualization: Stephan Schiffels.

Data curation: Iain Mathieson, Jared O'Connell.

Formal analysis: Ke Wang, Iain Mathieson, Stephan Schiffels.

Investigation: Ke Wang, Iain Mathieson.

Methodology: Ke Wang.

Project administration: Stephan Schiffels.

Resources: Jared O'Connell.

Software: Ke Wang.

Supervision: Stephan Schiffels.

Validation: Ke Wang.

Visualization: Ke Wang.

Writing – original draft: Ke Wang, Stephan Schiffels.

Writing – review & editing: Ke Wang, Iain Mathieson, Jared O'Connell, Stephan Schiffels.

References

1. McVean GAT, Cardin NJ. Approximating the coalescent with recombination. Philos Trans R Soc Lond B Biol Sci. 2005; 360: 1387–1393. https://doi.org/10.1098/rstb.2005.1673 PMID: 16048782
2. Marjoram P, Wall JD. Fast "coalescent" simulation. BMC Genet. 2006; 7: 16. https://doi.org/10.1186/1471-2156-7-16 PMID: 16539698
3. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011; 475: 493–496. https://doi.org/10.1038/nature10231 PMID: 21753753
4. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. 2014; 46: 919–925. https://doi.org/10.1038/ng.3015 PMID: 24952747
5. Steinrücken M, Kamm JA, Song YS. Inference of complex population histories using whole-genome sequences from multiple populations. Cold Spring Harbor Labs Journals; 2015 Sep. Available: http://biorxiv.org/lookup/doi/10.1101/026591
6. Sheehan S, Harris K, Song YS. Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. 2013; 194: 647–662. https://doi.org/10.1534/genetics.112.149096 PMID: 23608192
7. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet. 2017; 49: 303–309. https://doi.org/10.1038/ng.3748 PMID: 28024154
8. Kamm JA, Terhorst J, Song YS. Efficient computation of the joint sample frequency spectra for multiple populations. J Comput Graph Stat. 2017; 26: 182–194. https://doi.org/10.1080/10618600.2016.1159212 PMID: 28239248
9. Kamm J, Terhorst J, Durbin R, Song YS. Efficiently Inferring the Demographic History of Many Populations With Allele Count Data. J Am Stat Assoc. 2019; 1–16. https://doi.org/10.1080/01621459.2019.1635482
10. Excoffier L, Foll M. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics. 2011; 27: 1332–1334. https://doi.org/10.1093/bioinformatics/btr124 PMID: 21398675
11. Schiffels S, Haak W, Paajanen P, Llamas B, Popescu E, Loe L, et al. Iron Age and Anglo-Saxon genomes from East England reveal British migration history. Nat Commun. 2016; 7: 10408. https://doi.org/10.1038/ncomms10408 PMID: 26783965
12. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016; 538: 201–206. https://doi.org/10.1038/nature18964 PMID: 27654912
13. Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. A genomic history of Aboriginal Australia. Nature. 2016; 538: 207–214. https://doi.org/10.1038/nature18299 PMID: 27654914
14. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014; 505: 43–49. https://doi.org/10.1038/nature12886 PMID: 24352235
15. Plagnol V, Wall JD. Possible ancestral structure in human populations. PLoS Genet. 2006; 2: e105. https://doi.org/10.1371/journal.pgen.0020105 PMID: 16895447
16. Durvasula A, Sankararaman S. Recovering signals of ghost archaic admixture in the genomes of present-day Africans. bioRxiv. 2018. p. 285734. https://doi.org/10.1101/285734
17. Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, et al. Reconstructing Prehistoric African Population Structure. Cell. 2017; 171: 59–71.e21. https://doi.org/10.1016/j.cell.2017.08.049 PMID: 28938123

18. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The genomic landscape of Neanderthal ancestry in present-day humans. Nature. 2014; 507: 354–357. https://doi.org/10.1038/nature12961 PMID: 24476815

19. Sankararaman S, Mallick S, Patterson N, Reich D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. Curr Biol. 2016; 26: 1241–1247. https://doi.org/10.1016/j.cub.2016.03.037 PMID: 27032491

20. Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell. 2018; 173: 53–61.e9. https://doi.org/10.1016/j.cell.2018.02.031 PMID: 29551270

21. Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, et al. Genomic analyses inform on migration events during the peopling of Eurasia. Nature. 2016; 538: 238–242. https://doi.org/10.1038/nature19792 PMID: 27654910

22. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012; 338: 222–226. https://doi.org/10.1126/science.1224344 PMID: 22936568

23. Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature. 2014; 505: 87–91. https://doi.org/10.1038/nature12736 PMID: 24256729

24. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. 2013; 10: 5–6. https://doi.org/10.1038/nmeth.2307 PMID: 23269371

25. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007; 81: 1084–1097. https://doi.org/10.1086/521987 PMID: 17924348

26. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016; 48: 1443–1448. https://doi.org/10.1038/ng.3679 PMID: 27694958

27. Delaneau O, Howie B, Cox AJ, Zagury J-F, Marchini J. Haplotype estimation using sequencing reads. Am J Hum Genet. 2013; 93: 687–696. https://doi.org/10.1016/j.ajhg.2013.09.002 PMID: 24094745

28. Choi Y, Chan AP, Kirkness E, Telenti A, Schork NJ. Comparison of phasing strategies for whole human genomes. PLoS Genet. 2018; 14: e1007308. https://doi.org/10.1371/journal.pgen.1007308 PMID: 29621242

29. Song S, Sliwerska E, Emery S, Kidd JM. Modeling Human Population Separation History Using Physically Phased Genomes. Genetics. 2017; 205: 385–395. https://doi.org/10.1534/genetics.116.192963 PMID: 28049708

30. Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, et al. The genetic prehistory of southern Africa. Nat Commun. 2012; 3: 1143. https://doi.org/10.1038/ncomms2140 PMID: 23072811

31. Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, et al. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol Biol Evol. 2007; 24: 2180–2195. https://doi.org/10.1093/molbev/msm155 PMID: 17656633

32. Knight A, Underhill PA, Mortensen HM, Zhivotovsky LA, Lin AA, Henn BM, et al. African Y chromosome and mtDNA divergence provides insight into the history of click languages. Curr Biol. 2003; 13: 464–473. https://doi.org/10.1016/s0960-9822(03)00130-1 PMID: 12646128

33. Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science. 2012; 338: 374–379. https://doi.org/10.1126/science.1227721 PMID: 22997136

34. Schlebusch CM, Jakobsson M. Tales of Human Migration, Admixture, and Selection in Africa. Annu Rev Genomics Hum Genet. 2018. https://doi.org/10.1146/annurev-genom-083117-021759 PMID: 29727585

35. McDougall I, Brown FH, Fleagle JG. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature. 2005; 433: 733–736. https://doi.org/10.1038/nature03258 PMID: 15716951

36. White TD, Asfaw B, DeGusta D, Gilbert H, Richards GD, Suwa G, et al. Pleistocene Homo sapiens from Middle Awash, Ethiopia. Nature. 2003; 423: 742–747. https://doi.org/10.1038/nature01669 PMID: 12802332

37. Richter D, Grün R, Joannes-Boyau R, Steele TE, Amani F, Rué M, et al. The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age. Nature. 2017; 546: 293–296. https://doi.org/10.1038/nature22335 PMID: 28593967

38. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012; 192: 1065–1093. https://doi.org/10.1534/genetics.112.145037 PMID: 22960212

39. Hobolth A, Andersen LN, Mailund T. On computing the coalescence time density in an isolation-with migration model with few samples. Genetics. 2011. pp. 1241–1243.https://doi.org/10.1534/genetics.110.124164 PMID: 21321131

40. Kelleher J, Etheridge AM, McVean G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLoS Comput Biol. 2016; 12: e1004842. https://doi.org/10.1371/journal.pcbi.1004842 PMID: 27145223

5. Manuscript B

ANTHROPOLOGY

Ancient genomes reveal complex patterns of population movement, interaction, and replacement in sub-Saharan Africa

Ke Wang[1]*, Steven Goldstein[2]*, Madeleine Bleasdale[2], Bernard Clist[3,4], Koen Bostoen[3], Paul Bakwa-Lufu[5], Laura T. Buck[6,7], Alison Crowther[2,8], Alioune Dème[9], Roderick J. McIntosh[10], Julio Mercader[11,2], Christine Ogola[12], Robert C. Power[2,13], Elizabeth Sawchuk[2,14], Peter Robertshaw[15], Edwin N. Wilmsen[16,17], Michael Petraglia[2,8,18], Emmanuel Ndiema[12], Fredrick K. Manthi[12], Johannes Krause[1], Patrick Roberts[2,8], Nicole Boivin[2,8,11,18†], Stephan Schiffels[1†]

Africa hosts the greatest human genetic diversity globally, but legacies of ancient population interactions and dispersals across the continent remain understudied. Here, we report genome-wide data from 20 ancient sub-Saharan African individuals, including the first reported ancient DNA from the DRC, Uganda, and Botswana. These data demonstrate the contraction of diverse, once contiguous hunter-gatherer populations, and suggest the resistance to interaction with incoming pastoralists of delayed-return foragers in aquatic environments. We refine models for the spread of food producers into eastern and southern Africa, demonstrating more complex trajectories of admixture than previously suggested. In Botswana, we show that Bantu ancestry post-dates admixture between pastoralists and foragers, suggesting an earlier spread of pastoralism than farming to southern Africa. Our findings demonstrate how processes of migration and admixture have markedly reshaped the genetic map of sub-Saharan Africa in the past few millennia and highlight the utility of combined archaeological and archaeogenetic approaches.

INTRODUCTION

Africa today hosts enormous linguistic, cultural, and economic diversity. Reconstructing the patterns of population interaction, migration, admixture, and replacement that contributed to this diversity has been a core aim of genetic, archaeological, and linguistic studies for decades (*1*–*4*). As a relatively young field of research, ancient DNA (aDNA) has contributed less to these multidisciplinary efforts than other disciplines, and as a result of the limitations of skeletal and DNA preservation in Africa, aDNA has contributed less to African prehistory than elsewhere. While technical advances, such as the recognition of the petrous part of the temporal bone as a region that preserves high endogenous aDNA (*5*), have begun to change this situation, Africa remains understudied with only 85 ancient genomes published from the continent to date, relative to 3500 from Eurasia.

Previous aDNA studies from Africa have provided insights into population structure before the spread of food production in eastern and southern Africa (*2*, *3*, *6*) and revealed evidence for population turnovers in relation to changes in subsistence strategies in eastern Africa (*4*). Broadly, forager populations sampled between eastern and southern Africa were shown to have formed a continuous genetic cline roughly following geography (*3*). During the Pastoral Neolithic (PN), people related to Chalcolithic and Bronze Age Levantine groups entered eastern Africa and mixed there with individuals related to Later Stone Age foragers and with individuals related to present-day Dinka in what was proposed to have been at least a two-step process (*4*). Ancestry related to present-day Bantu speakers, which is, today, prevalent across sub-Saharan Africa, is absent from most ancient sub-Saharan African genomes analyzed to date.

Here, we report new insights into early population movements and admixture in Africa based on analysis of 20 newly generated ancient sub-Saharan African genomes (Table 1). Our sampling strategy follows a transregional approach to investigating population-level interactions between key groups that were identified previously as being involved in changes of food production strategies: eastern and southern forager groups, eastern African Pastoral Neolithic and Iron Age groups, and Iron Age groups related to present-day Bantu speakers. We sampled individuals from key regions where current models not only predict substantial interaction between foragers, herders, and farmers, particularly in eastern Africa, but also include the first individuals sampled from the Democratic Republic of the Congo (DRC), Botswana, and Uganda. By adding these new ancient genomes derived from archaeological forager and food-producing populations to published ancient and present-day sub-Saharan African genomes, we detect (i) evidence for the contraction of previously widespread and overlapping, deeply diverged forager populations; (ii) indications that the arrival of pastoral populations in eastern Africa resulted

[1]Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany. [2]Department of Archaeology, Max Planck Institute for the Science of Human History, Jena, Germany. [3]UGent Centre for Bantu Studies, Department of Languages and Cultures, Ghent University, Ghent, Belgium. [4]Institut des Mondes Africains, Paris, France. [5]Institut des Musées Nationaux du Congo, Kinshasa, Democratic Republic of Congo. [6]Department of Archaeology, University of Cambridge, Cambridge, UK. [7]Department of Anthropology, University of California, Davis, Davis, CA, USA. [8]School of Social Science, University of Queensland, St Lucia, Brisbane, QLD 4072, Australia. [9]Department of History, Cheikh Anta Diop University, Dakar, Senegal. [10]Department of Anthropology, Yale University, New Haven, CT, USA. [11]Department of Archaeology and Anthropology, University of Calgary, Calgary, Alberta, Canada. [12]Department of Earth Sciences, National Museums of Kenya, Nairobi, Kenya. [13]Institute for Pre- and Protohistoric Archaeology and Archaeology of the Roman Provinces, Ludwig-Maximilians-University Munich, Munich, Germany. [14]Department of Anthropology, Stony Brook University, Stony Brook, NY, USA. [15]Department of Anthropology, California State University, San Bernardino, San Bernardino, CA, USA. [16]University of Texas-Austin, Austin, TX, USA. [17]Witwatersrand University, Johannesburg, Republic of South Africa. [18]Department of Anthropology, Smithsonian Institution, Washington, DC, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: boivin@shh.mpg.de (N.B.); schiffels@shh.mpg.de (S.S.)

from the movement of several discrete groups of herders from northern to eastern Africa; and (iii) evidence for notable geographic diversity in patterns of herder-farmer-forager admixture during the spread of food production. These models are strengthened by integrating the first ancient genomes from the DRC, Botswana, and Uganda, allowing us to extend these multibranch models for the spread of food production across the continent. Data from Botswana also allow us to suggest a dispersal of eastern African pastoralists into southern Africa before the arrival of Bantu-speaking populations as has been previously suggested on the basis of linguistic and modern genetic data (*7*, *8*). Together, the ancient genomic and archaeological data examined here indicate that the economic heterogeneity that is the hallmark of modern Africa resulted from diverse local histories of population admixture, interaction, and avoidance.

RESULTS

New aDNA from Africa

We generated new genome-wide data from 20 ancient sub-Saharan African individuals (Table 1 and table S1), after screening skeletal material from 57 individuals (table S2). We evaluated the authenticity of aDNA for all screened samples based on characteristic cytosine-to-thymine deamination at the end of aDNA fragments and performed in-solution enrichment on mitochondria and 1.2 million autosomal single-nucleotide polymorphisms (SNPs) for 23 samples (two did not yield enough data after capture, and two samples were from the same individual) with endogenous DNA content above 0.1%. The successful samples include 5 individuals from the DRC [~795–200 before the present (BP)], 4 from Botswana (~1300–1000 BP), 1 from Uganda (~400–600 BP), and 10 from southern Kenya (~3900–300 BP), of which 3 are associated with eastern African foraging traditions, 5 with Pastoral Neolithic contexts, and 2 from the Iron Age. We combined these newly reported ancient genomes with previously published ancient African genomes (*2*–*4*, *9*–*11*), together with genomes from 584 individuals from 59 contemporary African populations (*1*, *12*), 44 high-coverage genomes from 22 African Indigenous populations (*13*), and 300 high-coverage genomes from 142 worldwide populations (*14*). The ages of the newly reported ancient individuals and their approximate sample locations are shown in Fig. 1. We examined the contamination level for all samples according to mitochondrial contamination estimates (*15*, *16*) and X chromosome contamination in males (table S1) (*17*). We also report mitochondrial haplogroups of each sample and Y chromosome haplogroups for most male samples (Table 1). We analyzed pairwise genetic similarities between all individuals and found that while NYA002 and NYA003 are consistent with being second-degree relatives, all other pairs are unrelated (see Materials and Methods).

Contraction of previously overlapping hunter-gatherer ancestries

We used principal components analysis (PCA) and model-based clustering to characterize the genetic relationship between our ancient individuals and published ancient and present-day African individuals (*1*–*4*, *9*–*14*). We find that our eight Kenyan samples, spanning 3900 to 1500 BP, form two clusters in PCA (Fig. 2), confirmed using ADMIXTURE (fig. S1) (*18*). Cluster 1 (named "east African foragers" in Fig. 2) consists of the new group/individual Kenya_Nyarindi_3500BP and Kenya_Kakapel_3900BP, as well as published data from Tanzania_Pemba_1400BP, Tanzania_Zanzibar_1400BP, and Kenya_400BP (Fig. 2). Cluster 2 (named "east African pastoralists") includes the new Kenyan samples with eastern African pastoralist-related ancestry. Individuals from cluster 1 show high genetic similarity to the 4500-BP hunter-gatherer from the Mota site in Ethiopia (*9*), as well as previously described ancient foragers from eastern Africa (*3*, *4*). We tested which ancestries other than Ethiopia_4500BP are present in these individuals although statistics of the form *f4* (ancient group, Ethiopia_4500BP; X, chimpanzee), which tests whether any other group X is more closely related to either our ancient individuals or Ethiopia_4500BP (the chimpanzee genome is required for technical reasons as an outgroup to all humans). Among the groups/individuals in this cluster (fig. S2), Kenya_Nyarindi_3500BP and Tanzania_Pemba_1400BP do not demonstrate significant genetic affinity to any other group that we tested here, while Kenya_Kakapel_3900BP shows significant genetic affinity with the Mbuti, a present-day group of Central African hunter-gatherers. In the same test, Tanzania_Zanzibar_1300BP has excess affinity with South_Africa_2000BP, as reported previously (*3*), and Kenya_400BP presents extra affinity with present-day west Eurasian people (*3*). We further characterized genetic ancestry components of these ancient African individuals through qpAdm (*19*), a method to estimate ancestry proportions related to specified source populations. We found Kenya_Kakapel_3900BP has 18 ± 6% Mbuti-related ancestry, and the published Kenya_400BP has 11 ± 3% ancestry related to ancient Levantine individuals (Fig. 3 and table S3), which likely reflects a gene pool present more broadly in ancient northeastern Africa and the Levant, as identified in ancient (*11*, *20*) and present-day northeastern African populations. These additional ancestral contributions are also seen on the PCA (Fig. 2) by their positioning relative to Ethiopia_4500BP. Modeling with qpAdm also suggests a small ancestry component related to southern African San in Kenya_Nyarindi_3500BP (models including San improve the fit significantly, but the resulting P value is still low, at $P = 0.002$).

Overall, these data point to eastern Africa as a nexus of population-level interactions between groups with ancestries associated with western, southern, and eastern African foragers. Deep divergences between these ancestries suggest either that admixture was minimal over a long period or that it occurred relatively recently. This poses interesting possibilities for more dynamic expansion and contraction of ancient African hunter-gatherer populations than have been postulated to date. Kenya_Kakapel_3900BP belongs to an archaeological fisher-forager group extending from Lake Victoria well into Uganda, and so the Mbuti-associated ancestry in this individual could be explained by ephemeral interactions between groups whose ranges overlapped when rainforest systems were more extensive in the early Holocene wet phase (*21*). Additional archaeological data from the region are needed to test this hypothesis.

Persistent detection of low levels of San-affiliated ancestry among ancient eastern African individuals is more difficult to explain. One possibility is ongoing interactions with an as-of-yet undetected hunter-gatherer population whose ancestry is primarily shared with the modern San. Another possibility is that the San-related ancestry reflects an earlier, wider distribution of African foragers stretching from southern to eastern Africa, which existed before Mid to Late Holocene migrations of farmers and herders (*3*). Linguistic and genetic parallels between eastern and southern African forager groups using click consonants make it tempting to hypothesize the presence of an early, widely distributed click language-speaking population (*1*, *22*), but there is no phylolinguistic evidence for a direct connection between these language groups (*23*).

Table 1. Summary of individuals with successful aDNA from Africa reported in this study. Note 1: Two samples from Lukenya Hill (LUK001 and LUK002) tend out to be genetically the same individual. We merged the genomic data for genetic analyses but report radiocarbon dates for both here. Note 2: The age of samples marked with asterisk is based on the archeological context instead of calibrated radiocarbon date. Aut. Cov., autosomal coverage; MT Cov., mitochondrial coverage.

ID	Population label	Archeological site	Country	Archeological affiliation	Aut. Cov.	MT Cov.	Sex	Y haplogroup	MT haplogroup	Aut. SNPs	SNPs hit on Human Origin dataset	Date (calendar BP)	Uncalibrated C14 dates ± error (lab number)
NYA002	Kenya_ Nyarindi_ 3500BP	Nyarindi Rockshelter	Kenya	Later Stone Age (Kansyore)	0.14	0.23	F	–	L4b2a	124,064	64,785	3555–3375	3253 ± 23 (OxA-37364)
NYA003	Kenya_ Nyarindi_ 3500BP	Nyarindi Rockshelter	Kenya	Later Stone Age (Kansyore)	0.02	0.02	M	E(E-M96,E-P162)	–	18,586	9736	–	–
LUK001	Kenya_ LukenyaHill_ 3500BP	Lukenya Hill, GvJm 202	Kenya	Pastoral Neolithic	0.59	1.41	M	E1b1b1b2b(E-M293,E-CTS10880)	L4b2a2b	495,472	222,439	3610–3460	3340 ± 23 (OxA-37356), 3296 ± 25 (OxA-37357)
LUK003	Kenya_ LukenyaHill_ 3500BP	Lukenya Hill, GvJm 202	Kenya	Pastoral Neolithic	0.01	0.30	F	–	L0f1	6830	3586	3635–3475	3359 ± 23 (OxA-37358)
HYR002	Kenya_ HyraxHill_ 2300BP	Hyrax Hill, GrJj25	Kenya	Pastoral Neolithic	0.77	0.52	M	E1b1b1b2b(E-M293,E-M293)	L5a1b	505,972	260,999	2365–2305	2354 ± 23 (OxA-37352)
MOL001	Kenya_ MoloCave_ 1500BP	Molo Cave, GoJi3	Kenya	Pastoral Neolithic	2.64	5.40	M	E1b1b1b2b(E-M293,E-M293)	L3h1a2a1	886,222	461,756	1415–1320	1532 ± 21 (OxA-37360)
MOL003	Kenya_ MoloCave_ 1500BP	Molo Cave, GoJi3	Kenya	Pastoral Neolithic	0.06	0.14	F	–	–	57,426	29,700	2110–1990	2101 ± 22 (OxA-37361)
KPL001	Kenya_ Kakapel_ 3900BP	Kakapel	Kenya	Later Stone Age (Kansyore)	0.92	3.94	M	CT(CT-M168,CT-M5695)	L3i1	572,074	299,181	3974–3831	3584 ± 28 [SUERC-86057 (GU51350)]
KPL002	Kenya_ Kakapel_ 300BP	Kakapel	Kenya	Later Iron Age/ protohistoric	1.26	78.35	F	–	L2a1f	684,698	363,447	309–145	222 ± 28 [SUERC-86058 (GU51351)]
KPL003	Kenya_ Kakapel_ 900BP	Kakapel	Kenya	Later Iron Age	0.07	63.21	F	–	L2a5	75,113	39,367	910–736	895 ± 28 [SUERC-86059 (GU51352)]
MUN001*	Uganda_ Munsa_ 500BP	Munsa	Uganda	Later Iron Age	0.46	1.57	F	–	L3b1a1	377,332	–	1400–1600 CE	–
KIN002	Congo_ Kindoki_ 230BP	Kindoki	DR Congo	Protohistoric	0.62	1.46	M	E1b1a1a1d1a2(E-CTS99,E-CTS99)	L1c3a1b	438,125	229,240	295–145	217 ± 20 (OxA-37353)
KIN003	Congo_ Kindoki_ 150BP	Kindoki	DR Congo	Protohistoric	0.02	0.09	M	E(E-M96,E-PF1620)	–	19,691	10,329	285–modern	172 ± 20 (OxA-37354)

ID	Population label	Archeological site	Country	Archeological affiliation	Aut. Cov.	MT Cov.	Sex	Y haplogroup	MT haplogroup	Aut. SNPs	SNPs hit on Human Origin dataset	Date (calendar BP)	Uncalibrated C14 dates ± error (lab number)
KIN004	Congo_ Kindoki_230BP	Kindoki	DR Congo	Protohistoric	0.96	2.01	M	R1b1(R-P25_1,R-M415)	L0a1b1a1	560,376	291,465	305–150	241 ± 20 (OxA-37355)
NGO001	Congo_ NgongoMbata_ 220BP	Ngongo Mbata	DR Congo	Protohistoric	0.42	0.78	M	–	L1c3a	328,389	170,742	295–145	211 ± 21 (OxA-37363)
MTN001	Congo_ MatangaiTuru_ 750BP	Matangai Turu Northwest	DR Congo	Iron Age forager	0.06	0.33	F	–	–	52,012	28,452	795–690	871 ± 21 (OxA-37362)
NQO002*	Botswana_ Nqoma_ 900BP	Nqoma	Botswana	Early Iron Age	0.02	0.60	F	–	L2a1f	14,189	7,587	700-1090 CE	–
TAU001*	Botswana_ Taukome_ 1100BP	Taukome	Botswana	Early Iron Age	0.09	5.82	M	E1b1a1(E-M2,E-Z1123)	L0d3b1	79,261	42,998	900-1000 CE	–
XAR001*	Botswana_ Xaro_ 1400BP	Xaro	Botswana	Early Iron Age	3.64	37.94	M	E1b1a1a1c1a	L3e1a2	939,378	494,074	700-1000 CE	–
XAR002*	Botswana_ Xaro_ 1400BP	Xaro	Botswana	Early Iron Age	1.36	172.94	M	E1b1b1b2b(E-M293,E-CTS10880)	L0k1a2	703,295	375,283	700-1000 CE	–

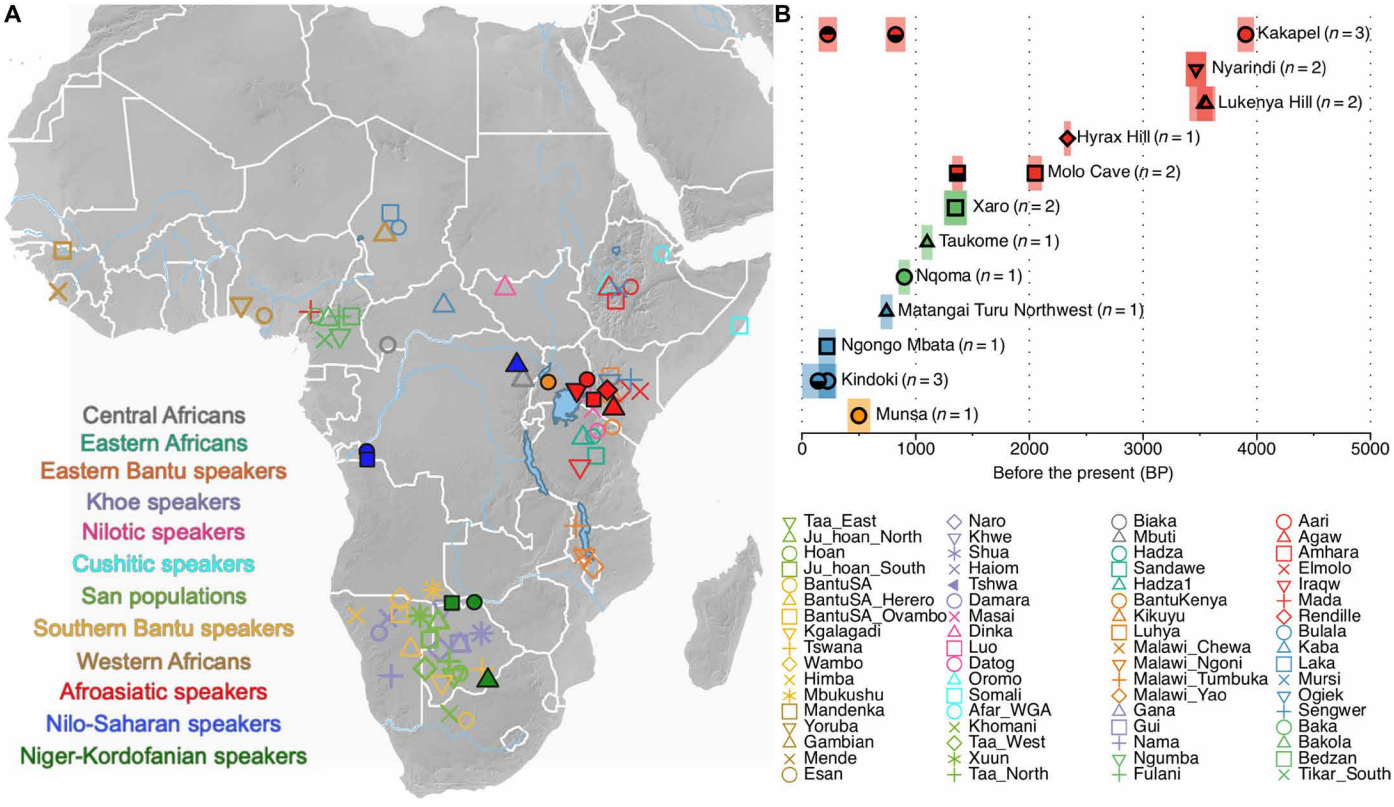


Fig. 1. Basic information of newly reported ancient genomes. (**A**) Approximate locations of new samples and published present-day modern African populations. Same legend scheme applies to the principal components analysis (PCA) plot in Fig. 2. (**B**) C14 dates after calibration. Samples from Botswana (green) and Uganda (orange) are based on archaeological context dates rather than accelerator mass spectrometry (AMS) measurements.

Complex spread of pastoralism to eastern Africa

Cluster 2 of the Kenyan samples on the PCA (Fig. 2), with east African pastoralist-related ancestry, includes the newly reported groups/individuals from sites of the Savanna Pastoral Neolithic tradition in South Kenya: Kenya_LukenyaHill_3500BP, Kenya_HyraxHill_2300BP, and Kenya_MoloCave_1500BP, which fall into the beginning, middle, and end, respectively, of the Pastoral Neolithic period in Kenya, as well as a published ancient genome from Tanzania, Tanzania_Luxmanda_3100BP (*3*), and other published Pastoral Neolithic genomes from eastern Africa (*4*). These samples show remarkable continuity of ancestry across a time span of 2000 years, presenting similar genetic profiles in PCA and clustering analysis (Fig. 2 and fig. S1).

On the basis of previous models for Tanzania_Luxmanda_3100BP (*3*), we first applied two-way ancestry models in qpAdm using Ethiopia_4500BP and a group of ancient Levantine individuals (*24*), which we take as the closest available proxy for ancient northeastern African ancestry (*10*, *11*), as sources. Consistent with the findings of a previous aDNA study (*4*), we found this model to be insufficient (Fig. 3 and table S3) and demonstrate that an additional genetic component related to the present-day Dinka (a Nilotic-speaking group from South Sudan) is necessary to fit the data. In addition to qpAdm, we confirmed this affinity using a customized *f4* test (see Materials and Methods and fig. S3). In our final three-way model, which is qualitatively similar to the model proposed in (*4*), we find 33 ± 11% and 24 ± 10% Dinka-related ancestry in Kenya_HyraxHill_2300BP and Kenya_LukenyaHill_3500BP, respectively, and lower proportions in Kenya_MoloCave_1500BP and Tanzania_Lxumanda_3100BP (Fig. 3 and table S3).

While the estimated proportions of Levantine-related ancestry in all samples are rather constant (around 30 to 40%), we find that both the proportion of east African forager-related ancestry, as well as of Dinka-related ancestry, varies substantially across individuals. An earlier study (*4*) concluded that admixture between pioneering herders with Levantine-related ancestry and eastern African hunter-gatherers primarily occurred before their arrival in southern Kenya. However, our data suggest that periodic admixture between herders and hunter-gatherers, or populations predominantly carrying ancestry derived from them, may have continued into the PN. In particular, the newly reported 1500-BP individuals from Molo Cave carry 50% or more forager-related ancestry, and less Dinka-related ancestry, than observed in all other sequenced Pastoral Neolithic individuals (*4*). A model of repeated interaction between foragers and herders is further supported by admixture date estimates using linkage disequilibrium decay, which suggest that admixture dates between ancestry related to Chalcolithic Levant (*24*) and to Ethiopia_4500BP range from a few hundred to a few thousand years before the time of death of the individuals, with no clear correlation between admixture age and age of sample (fig. S4), inconsistent with a simple model of admixture, but suggesting either multiple events, or strong population structure preventing homogenization of ancestries over a long time period. Despite only minimal archaeological evidence for the persistence of autochthonous hunter-gatherers in the Central Rift Valley this late in the PN (*25*), these genetic results suggest

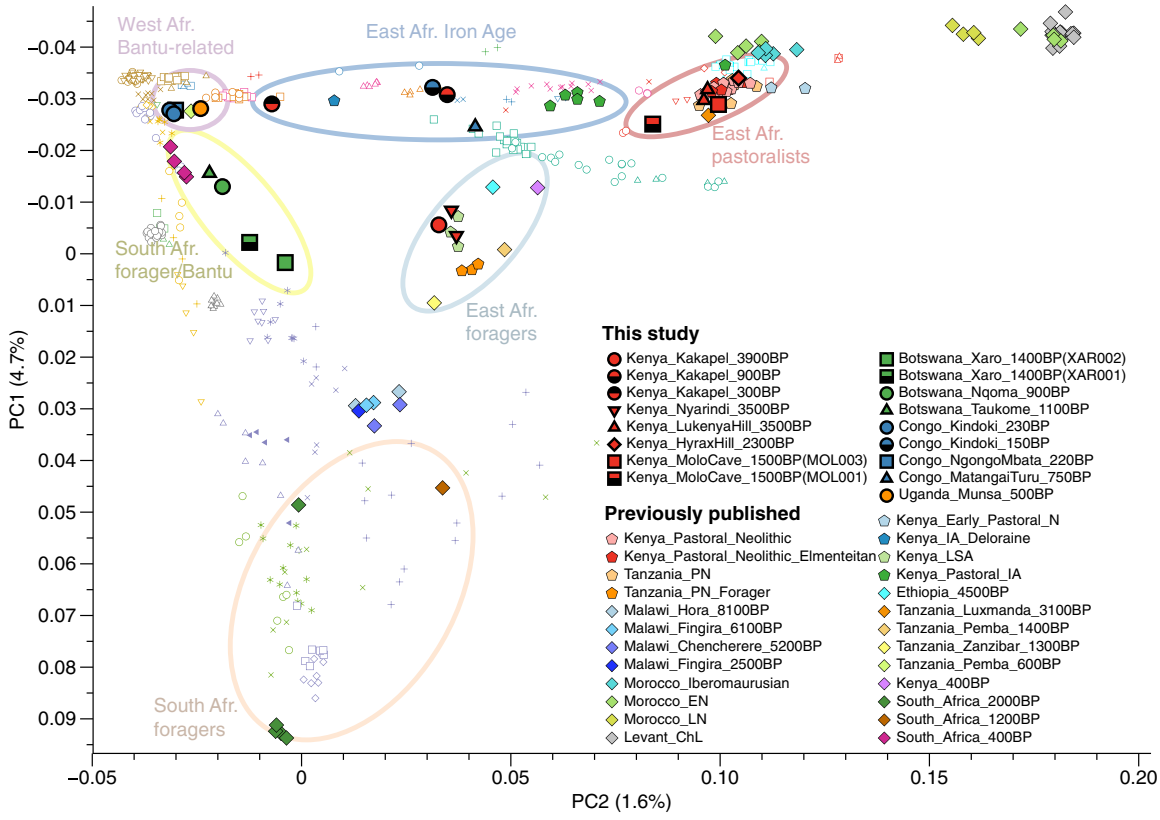


Fig. 2. PCA of ancient genomic data analyzed in this and previous studies, together with published modern genetic data. Modern populations shown are detailed in the legend of Fig. 1 and fig. S1. A separate PCA with only present-day populations is shown in fig. S1A. Color circles highlight the key groups discussed in this paper and summarized in Fig. 3A.

that communities with high or unadmixed hunter-gatherer–related ancestry continued to live alongside communities with high or unadmixed Pastoral-Neolithic related ancestry until nearly the Iron Age, leaving prominent genetic traces at Molo Cave. It is not yet clear from Molo Cave or other sites whether the timing and pace of admixture reflects adoption of herding by foragers, absorption of foragers into herding groups, or more complex intergroup social dynamics.

Combining evidence from both eastern African genetic clusters, we document very different patterns of interaction and admixture from sampled individuals along the eastern African and Lake Victoria shores relative to the patterns in the Central Rift. Near lake and ocean coasts, we see little evidence for pastoralist admixture into forager individuals [e.g., Kenya_Nyarindi_3500BP and two previously sampled individuals from Zanzibar (*3*)]. Our analysis also demonstrates that the recently published individual from the cave site of Panga ya Saidi in coastal Kenya [Kenya_400BP (*3*)] similarly retains a predominantly eastern African forager ancestry, with only a small Levantine-related component. This is the exact opposite of the pattern observed in individuals around the Central Rift, where pastoralist-mediated, Levantine-related ancestry spread rapidly. It may be that delayed-return foragers in stable coastal and lacustrine environments were more demographically numerous and/or resistant to interactions with incoming food producers than other hunter-gatherers.

While our data support the three-component model for the Pastoral Neolithic (*4*), our findings suggest greater complexity than initially proposed for the admixture of existing and incoming populations in this period. The fact that both Dinka-related ancestry and eastern African forager-related ancestry varies substantially in our samples and previously published samples suggests that the spread of herding either involved complex population structure maintained over a long time period or prevented homogenization of these ancestries, or multiple population movements with regionally distinct trajectories of interaction and admixture. This adds increasing resolution to proposed diversity of populations that contributed to the "moving frontier" model for herder dispersals in eastern Africa (*4*, *26*). Individuals from Molo Cave, Luxmanda, and Panga ya Saidi furthermore provide evidence that contact with eastern African foragers, who coexisted with food producing people until at least 400 BP (Fig. 3A), was a continuous process, rather than one that occurred only during initial phases of contact.

The data also reveal that this interaction between herders and foragers was very imbalanced, with hunter-gatherer ancestry entering pastoralist populations, but little flow in the other direction. It is not clear what forms of social systems between herders and foragers may have resulted in this one-way admixture. In the past, it has been assumed that low herder population density and high risk of herd loss from epizootic disease would require herders to form closer relationships with local hunter-gatherers who had greater ecological knowledge of the landscape (*27*, *28*). This has been supported by evidence for herder-forager interactions at sites such as Crescent Island (*29*) and Prolonged Drift (*30*). Genetic evidence indicates that if these interactions occurred, then they were more structured and possibly

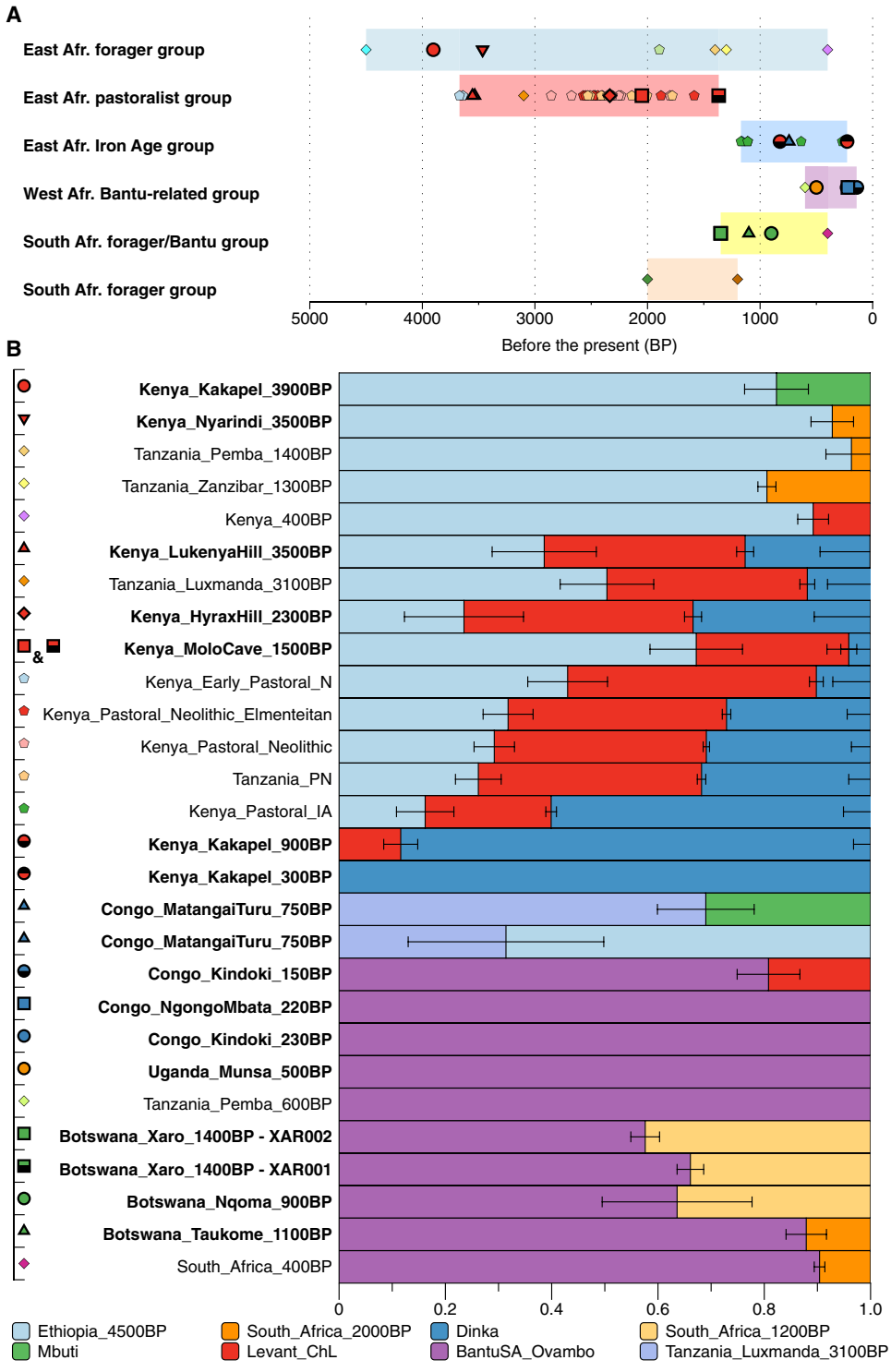


Fig. 3. Admixture history of ancient African populations. (**A**) Overview of coexistence of distinct African ancestries through time drawing on currently available ancient genomes. (**B**) Ancestral components of ancient African groups/individuals according to qpAdm. We order ancient groups in the same order shown in (A) and highlight newly reported genetic groups/individuals in bold. *P* values and estimated ancestral proportions can be found in table S3.

more consistent with ethnographic client-patron relationships (*31*), wherein individuals from hunter-gatherer communities may be slowly integrated into herder societies. It is possible that sex bias due to different social dynamics played a role in the observed asymmetric gene flow between the two groups. While we could not test this explicitly due to insufficient coverage on the X chromosomes, these dynamics have been previously described between foragers in central and southern Africa and Bantu-speaking farmers (*32*).

Shifts of ancestry during the Iron Age in central and eastern Africa

Three new samples also allowed us to evaluate changes in ancestry during the Iron Age. The Kakapel site in western Kenya, from which we analyzed the 3900-year-old forager above, also featured two Iron Age individuals (Kenya_Kakapel_300BP and Kenya_Kakapel_900BP), which show close genetic affinity to Dinka and other Nilotic-speaking groups (Luo, Datog, and Maasai) using PCA and ADMIXTURE, and also have closer genetic affinity with present-day Bantu speakers than ancient foragers or Pastoral Neolithic individuals (Fig. 2 and figs. S1 and S5).

On the basis of the affinity seen on the PCA, we tested whether Kenya_Kakapel_300BP and Kenya_Kakapel_900BP are genetically similar to the Nilotic-speaking Dinka and Luo and Bantu-speaking Luhya and Kikuyu (all are ethnic groups in modern Kenya, except the Dinka of South Sudan). Using *f4* statistics and qpAdm, we find that Kenya_Kakapel_300BP is similar in ancestry to Dinka, with Luo and Luhya providing marginally fitting models as well (Fig. 3, fig. S5, and table S4). Kenya_Kakapel_900BP also shares close genetic affinity with Dinka but requires an additional small ancestry component (12 ± 3%) from northeastern African/Levantine groups, similar to the ancestry component in early PN herders (Fig. 3 and table S4). We dated this admixture between Dinka- and Levantine-related ancestries in Kenya_Kakapel_900BP to around 500 ± 200 years before the death of that individual, consistent with the onset of the Iron Age in the region. This suggests that the Iron Age population represented by this single individual resulted from admixture between PN-related herders and incoming Nilotic agropastoralists, rather than resulting from a major migration of people with West African–related ancestries.

The notable shift seen in the two Iron Age individuals from the Kakapel site to almost 90 to 100% Nilotic-related ancestry, compared to about 40% during the Pastoral Neolithic, is substantially larger than the increase in Nilotic ancestry seen in previously analyzed eastern African individuals from the Iron Age (*4*). In addition, the absence of ancestry related to present-day Bantu speakers in Kenya_Kakapel_900BP contrasts with the finding of this ancestry in a contemporaneous individual from the site of Deloraine farm in the Central Rift Valley of Kenya (*4*). This shows that patterns of dispersal and admixture in Iron Age eastern Africa resulted in a complex geography of ancestry, with some regions or locations witnessing almost complete replacement from Nilotic-related migrations (*33*), others seeing mixing of diverse peoples (*4*), and yet others demonstrating no admixture from ancestry related to Nilotic or Bantu speakers into recent centuries (as seen in Kenya_400BP).

Previous research associated the increase in Nilotic ancestry during the Iron Age with a so-called "Pastoral Iron Age" based on samples from the Central Rift Valley (*4*). Our findings for the Iron Age, much like our findings for the PN, are consistent with multiple groups with different subsistence systems entering eastern Africa along different geographical routes. While these can broadly be grouped as a single "stage" of population change (*4*), it is increasingly clear that there is greater heterogeneity in the nature of population change within southern Kenya than previously recognized.

A new Iron Age genome from the eastern border of the DRC (Congo_MatangaiTuru_750BP) highlights additional trajectories of forager–food producer interaction as herding and farming spread into Central Africa. The best-fitting model for this individual is one including Ethiopia_4500BP as one source and Pastoral Neolithic as the other (Fig. 3 and table S5). We tested an alternative model with Mbuti instead of Ethiopia_4500BP, which also provided a working fit and which fits a signal seen on PCA (specifically, PC4; see fig. S8), which shows that this individual is shifted toward Mbuti. While the sparse genetic data available for the Matangai Turu individual did not allow us to select between these two models, we highlight that both models indicate PN-related ancestry in a region hitherto unsampled for aDNA. We argue that this finding may reflect continued expansion of Pastoral Neolithic populations, with or without herding, during the Iron Age, possibly related, or in response to, displacement by incoming groups related to Nilotic- and Bantu-speaking populations. We caution that this argument is based on a single individual and more data from the region are necessary to make stronger statements. Our successful aDNA extraction from a rainforest location shows that this is possible.

A single sample from Munsa, Uganda, indirectly dated to the 14th to 16th century CE (*34*), together with the published Tanzania_Pemba_600BP individual, documents the dispersal of ancestry related to present-day Bantu speakers throughout eastern Africa (Fig. 3 and table S5). This individual likely also reflects a Bantu-speaking population in Uganda during a period of complex-state formation in association with cattle keeping and cereal cultivation (*34*).

Direct evidence of genetic exchange between Bantu and pastoralist/foragers in southern Africa

New ancient genomes from Botswana (three ancient individuals from the Okavango Delta region of northwestern Botswana and one from southeastern Botswana) allowed us to extend investigation of the spread of food-producing populations into southern Africa. Positioning on the PCA suggests mostly ancestry related to present-day Bantu speakers in these individuals (Fig. 2), and our modeling shows that the dominant genetic ancestry component in all four Botswana individuals is related to BantuSA_Ovambo, the Bantu-speaking southern African Ovambo (Fig. 3). Given the geographic position of the individuals and the genetic position on the PCA, we suspected another genetic ancestry component related to southern African hunter-gatherers. We therefore tested both South_Africa_2000BP and South_Africa_1200BP in two-way models for all four individuals (Figs. 2 and 3 and table S6). This provided working models for all individuals, with 30 to 40% southern African hunter-gatherer ancestry for the three individuals from the Okavango Delta (Nqoma and Xaro) and around 10% for the individual from the eastern border of Botswana (Taukome).

While, for the Nqoma and Taukome individuals, both southern African sources fit the data, for the two Xaro individuals, only South_Africa_1200BP provides a working fit, while South_Africa_2000BP fails (table S6). While South_Africa_2000BP has unadmixed southern African hunter-gatherer ancestry, South_Africa_1200BP was shown to be admixed with eastern PN-related ancestry (*3*), a pattern present in most Khoisan groups today (*1*). The fact that only South_Africa_1200BP provides a fitting model for the two Xaro individuals therefore suggests PN-related ancestry in these individuals, and we argue that our findings point to the presence of the same ancestry in the third individual from the Okavango Delta (Nqoma), although the low coverage in that individual prevents us from testing this. We also assessed whether the ancient Botswana individuals have differential ancestry to present-day Khoisan groups and found that only Juhoan_North stands out in that it has less affinity to ancient Botswana individuals compared to Gui, Naro or Juhoan_South (see table S9). An assessment using different Bantu sources in our qpAdm modeling shows

that among different proxies of ancestry related to present-day Bantu speakers, only BantuSA_Ovambo, a group of southwestern Bantu speaker from Namibia, provides working models, while Tswana and Kgalagdi, who are most populations of Botswana and among southeastern Bantu speakers today, failed in our statistical modeling (table S6).

We confirmed PN-related ancestry by fitting three-way models with the Pastoral Neolithic individual Tanzania_Luxmanda_3100BP and South_Africa_2000BP as additional sources on top of BantuSA_Ovambo. Botswana_Xaro_1400BP and Botswana_Nqoma_900BP show 14 to 22% ancestral contribution from the PN source. Consistently, uniparental markers in the two individuals from Xaro support mixed ancestry. The first individual (XAR001) has mitochondrial haplogroup L3e1a2 and Y chromosome haplogroup E1b1a1a1c1a, both common in Bantu-speaking populations (*35*, *36*). The second individual (XAR002) has Y haplogroup E1b1b1b2b, associated with most ancient eastern African pastoralists analyzed here and previously (fig. S9), and also found in present-day southern African pastoralists (*37*), while his maternal lineage (L0k1a2) is possibly of indigenous South African Khoisan origin (*36*).

We assessed which ancestry (related to Neolithic pastoralists or Bantu speakers) admixed first with the South African forager-related gene pool using linkage disequilibrium decay (fig. S7) and could show that eastern African pastoralist-related admixture generally predates admixture from ancestry related to Bantu speakers. This is consistent with previous models of South African population history based on modern African genomes (*1*) and with linguistic (*7*) and archaeological (*38*) hypotheses for eastern African herders becoming established in this region before the Iron Age. We emphasize that our data do not address where the mixture between eastern herders and southern hunter-gatherers occurred. However, the aDNA data clearly point to the presence of already admixed southern forager and eastern pastoralist ancestry in the Okavango Delta by the late first millennium CE (Fig. 3). The order of admixture events in Botswana is directly supported by the ancestry mix present in the Okavango Delta individuals from a Bantu-related source and a South_Africa_1200BP-related source. Conversely, if admixture between ancestors of Bantu-speaking and eastern African herder populations had occurred before input of southern hunter-gatherer ancestry in southern Africa, then these signatures would be apparent in other regions, but, so far, early arrivals of Bantu speakers in nearby Malawi do not carry this eastern African component (*3*). Rather, in the most parsimonious model, initial population mixture occurred between groups related to South_Africa_2000BP and eastern African pastoralists (with South_Africa_1200BP being a descendant of that initial mixture). Bantu speakers arriving in southern Africa then mixed with this population giving rise to the individuals from Xaro analyzed here. No present-day population sampled so far has the same ancestry mix as the two Xaro individuals (as visible from the PCA; Fig. 2). While further sampling may still reveal such a population in the future, so far, this suggests that this population was later replaced by unadmixed Bantu-speaking populations, as inhabit the region today.

The arrival of East-African pastoralist-related ancestry in Botswana and South Africa has been associated with the emergence of lactase persistence (LP) in these regions, as found in some Khoe-speaking people today, such as the Nama (*39*, *40*). We therefore investigated whether any of the known SNP alleles associated with LP are present in the ancient Botswana individuals or any of the other African individuals reported in this study. Among eight LP-related SNP positions that are present in our 1240K capture panel, we found no evidence for the presence of any of these LP-associated alleles (table S7). We also examined malaria resistance genes, which have been linked to the spread of Bantu speakers, and found derived alleles in XAR002 at SNPs rs2515904 and rs1050829 (table S7), where derived alleles are associated with a higher risk to malaria (*41*), coinciding with the admixture with ancestry related to Bantu speakers found in the genetic profile of this individual (table S6).

Genetic results mirror archaeological data indicating diversity in the emphasis on farming, herding, and foraging between sites and communities during the early Iron Age of Botswana (*42*, *43*). As in eastern Africa, it appears that specific trajectories of interaction and integration in particular regional and temporal settings influenced the diversity in subsistence strategies that was a hallmark of African history until recent centuries.

Historical individuals from Congo document ancestry related to Bantu speakers in Central Africa

Our most recent ancient genomes come from the west of the DRC (Congo_Kindoki_230BP and Congo_NgongoMbata_220BP) and show unadmixed ancestry related to present-day Bantu speakers, similar to the individual from Munsa analyzed above, clustering tightly together in the PCA with the published individual Tanzania_Pemba_600BP and some present-day eastern and southern African Bantu speakers (Fig. 2). Grouping Congo_Kindoki_230BP and Congo_NgongoMbata_220BP as a single genetic group, we tested their genetic affinity to present-day Bantu-speaking populations and ancient genomes related to present-day Bantu speakers, including Munsa, via outgroup *f3* statistics. Our samples share highest genetic affinity with the ancient individuals Tanzania_Pemba_600BP and Kenya_IA_Deloraine, followed by BantuSA_Ovambo. We further found no other population that has more genetic affinity to either the ancient Congo individuals or BantuSA_Ovambo than Tanzania_Pemba_600BP, using the symmetry test *f4* (Congo_Kindoki_NgongoMbata, BantuSA_Ovambo; X, chimpanzee) (fig. S6), which is also confirmed by qpWave (table S5). The fact that the ancient individuals with ancestry related to Bantu speakers are more closely related to each other than to present-day Bantu-speaking groups, despite the notable temporal and spatial distance between them, might reflect input of additional ancestral components in most present-day Bantu-speaking populations as a result of later migrations but could also be confounded by batch effects among aDNA samples being generally slightly attracted to each other compared to present-day genotyping data. It should also be noted that evident gaps in the sampling of present-day populations exist, including in the DRC itself and many neighboring countries.

The other ancient individual from Kindoki, Congo_Kindoki_150BP, presents a genetic makeup different from Congo_Kindoki_230BP, based on PCA and admixture analysis (Fig. 2). Again, grouping Congo_Kindoki_230BP with Congo_NgongoMbata_220BP, we performed *f4* statistics for testing whether Congo_Kindoki_150BP and the two other historic groups are genetically similar. As shown (fig. S6D), several west Eurasian groups (or ancient African groups carrying west Eurasian ancestry) are genetically significantly closer to Congo_Kindoki_150BP than to the other Congo individuals. When modeling Congo_Kindoki_150BP with qpAdm (Fig. 3B and table S5), we found a fitting model with 85 ± 7% ancestry related to Bantu speakers and 15 ± 7% ancestry related to western Eurasians. This ancestry profile would be consistent with the hypothesis that this individual

has Portuguese ancestry, which would fit with the colonial history of the region (*44*) and the Christian burial of this and other individuals in Kindoki (see Supplementary Text).

DISCUSSION

Our study documents the coexistence, mobility, interaction, and admixture of diverse human groups throughout sub-Saharan Africa over the past few thousand years by describing 20 new ancient genomes from Kenya, Uganda, the DRC, and Botswana. Together with previously published ancient African genomes (*3*, *4*, *9*), it demonstrates that, across all regions studied, the earliest visible ancestry is closely related to that of present-day hunter-gatherer populations such as the San in southern Africa, the Hadza in eastern Africa, and the Mbuti of the central African rainforest. Current data show that while this geographically defined forager population structure extends back to at least the mid-Holocene in eastern Africa (as represented by the 4500-BP individual from Mota), current forager populations reflect a contraction of ancestries that were once more spatially overlapping [as noted in (*3*) for eastern and southern hunter-gatherers]. Restriction of gene flow between regional forager groups in eastern, southern, and central Africa, whether over the long term due to climatic and environmental factors such as increasing aridity or later as a result of encapsulation by food-producing groups, has likely contributed significantly to the population structure observed in the African continent.

It is worth noting that, in some cases, overlapping forager ancestries could also reflect prefood-producing era migrations. For example, it is possible that the expansion of bone harpoon technologies (*45*), wavy-line pottery (*46*), and aquatic resource-based economies from northern to eastern Africa in the early Holocene also involved population migrations (*21*). The wetter climate conditions at the time may also have encouraged previously invisible east-to-west connections between hunter-gatherers in the central African rainforests and the eastern African Great Lakes, perhaps reflected in the Mbuti-related ancestry in our early sample from Kakapel.

Our six new individuals from the Pastoral Neolithic in Kenya were added to previous findings (*4*), demonstrating greater complexity in their ancestry profiles than previously observed for Pastoral Neolithic individuals from the same region (*4*). While this may be the result of population structure preventing random mating and homogenization, another explanation for this pattern is that early herders migrated south along multiple contemporaneous, but geographically distinct, routes in a manner similar to historic branching migrations of Maa, Ateker, and Surmic peoples across eastern Africa. In such a scenario, a single-base population in northern Africa may have branched into many as some herding groups moved along the Nile corridor, some through southern Ethiopia, and possibly some through eastern Uganda. Following varying trajectories, groups would have encountered different populations and formed diverse patterns of intercommunity relationships, resulting in more variable integration of ancestries. This model may help explain why stark variations in material culture, settlement strategies, and burial traditions are maintained for so long among PN populations with closely shared ancestries. Furthermore, detection of substantial eastern African forager ancestries late in the PN at Molo Cave indicates a longer persistence of indigenous foragers than is evident in the archaeological record (*25*). Despite appearing genetically homogenous overall, forager groups interacted with incoming herders with different degrees of resistance or integration (*27*) that affected the timing and structure of genetic admixture. Additional archaeological and archaeogenetic data are still needed to test this model and better reconstruct historically contingent patterns of migration and interaction.

Moving into the Iron Age, we again see evidence for multiple pathways of population movement in eastern, central, and southern Africa. The two Iron Age individuals from the Kakapel site near Lake Victoria document a more extreme (and near-complete) increase in Nilotic-related ancestry, possibly related to the arrival of the Luo, than the five previously published Iron Age individuals from the Central Rift Valley (*4*). The only explanation for this is that genetic turnover must have been region-specific and could have involved multiple divergent migrations. Our observation of PN-related ancestry in eastern Congo in the late Iron Age, as well as the lack of ancestry related to Bantu speakers there at that time, is, so far, an isolated find that calls for further investigations about the spread of PN-related ancestry in the west of the eastern African core region.

The interplay between incoming Bantu speakers (as evidenced by ancestry in present-day groups such as the Luhya and Kikuyu) and Iron Age Dinka-related ancestry remains unclear, including the question of whether farming spread exclusively through the expansion of Bantu-speaking populations, or also through local adoption (*47*). However, new ancient genomic data from this study track the footprint of migrating Bantu speakers further into the south. Our data document the arrival of people with ancestry related to Bantu speakers in Botswana in the first millennium CE and their admixture there with eastern African pastoralist and southern African forager ancestry. It provides evidence for interactions between three distinct lineages in the region, in line with the hypothesized arrival of Bantu-speaking communities into southern Africa by 1700 BP (*48*), and offers genetic support to the hypothesis of a pre-Bantu expansion of pastoralists into southern Africa (*3*, *7*, *38*).

Beyond the signature of ancestry related to Bantu speakers in southern Africa, we also find this ancestry in unadmixed form in historical individuals from Uganda and western Congo, which show a genetic profile similar to that of previously published individuals from Tanzania [Tanzania_Pemba_600BP (*3*)] and Deloraine Farm [Kenya_IA_Deloraine (*4*)], as well as present-day Southern Bantu speakers (BantuSA_Ovambo), consistent with the well-documented genetic homogenization caused by the Bantu expansion (*49*). Nonetheless, aDNA studies are beginning to reveal highly variable patterns of Bantu admixture with regional forager and pastoralist populations in sub-Saharan Africa, with unadmixed ancestry related to Bantu speakers persisting in the western Congo and Tanzania until the historical era, but evidence for noticeable admixture within centuries of initial arrival of Bantu speakers in southern Africa (*3*).

Our study highlights that while supraregional studies such as this one are important to understand continental-scale processes, increasingly regional-focused studies are called for in the future to better understand region-specific patterns of cultural and population changes (*4*). Important focal regions for these studies would include Sudan and the Horn of Africa to better understand the processes that brought the first herders into eastern Africa and regions to the north of Botswana, such as Zambia, to reveal more details about the interactions between early pastoralists and South African hunter-gatherers, as revealed by our individuals from Botswana. These studies are becoming more and more possible given the promising and increasing success rate of aDNA from Africa in a diversity of settings and time periods.

MATERIALS AND METHODS

Material collection

All sampling material from Kenya was sampled and exported under permits issued by the National Museums of Kenya and permissions from the National Commission for Science, Technology, and Innovation, Kenya. Material from Uganda was exported under a Ugandan government permit. Material from the DRC was excavated, sampled, and exported as part of the KongoKing project as outlined in text S1. The material from Botswana was exported under available permits from the Botswana government.

Direct accelerator mass spectrometry ^{14}C bone dates

We report 15 new direct accelerator mass spectrometry (AMS) ^{14}C bone dates in this study from two radiocarbon laboratories (Oxford, 13; Glasgow, 3). Bone samples were prepared following the laboratory-specific protocol for radiocarbon dating. All ^{14}C ages were calibrated with the IntCal13 Northern Hemisphere calibration curve (*50*) using OxCal version 4.3.2 (*51*). All uncalibrated, calibrated, and context-based dates are summarized in Table 1.

aDNA sample processing

Originally, we screened 56 skeletal samples for DNA preservation from seven collections from different institutions (table S2) in dedicated clean rooms at the Max Planck Institute for the Science of Human History in Jena, Germany. DNA extraction and library preparation were performed with previously published protocols (*52*), including partial uracil-DNA glycosylase treatment (*53*) to reduce the characteristic deamination error of aDNA fragment. After screening, we enriched for 1.2 million informative nuclear SNPs (1240K) by in-solution hybridization (*54*) for 20 samples with ≥0.1% endogenous content. We processed DNA sequences using the EAGER v1.92.50 pipeline (*55*), with adaptors removed by AdapterRemoval v2 (*56*), reads mapped to hs37d5 by BWA alignment software v0.7.12 (*57*), and polymerase chain reaction duplicates removed by Dedup software v0.12.2 (*55*). We trimmed the first and last 3 base pairs (bp) of each read using *trimBam* function in bamUtils v1.0.13 (*58*). We applied a minimum base quality (Phred-scaled) of 30 and a minimum mapping quality (Phred-scaled) of 30- to 3-bp masked BAM files for contamination estimates and calling genotypes. We called a random allele for each target SNP after quality-filtering to produce a pseudo-diploid genotype. For most of the downstream population genetic analyses, we used all autosomal SNPs from 1240K capture, while for a subset of analyses, we used transversions only to avoid the aDNA deamination error at transition sites. Mitochondrial DNA contamination was estimated using Schmutzi (*15*). For males, we estimated nuclear contamination using ANGSD v0.910 (*17*). All contamination estimates can be found in table S1.

Uniparental haplogroup and kinship analysis

For mitochondrial DNA haplogroups, we used HaploGrep2 (*59*) and HaploFind (*60*) with mitochondrial consensus sequences generated by Geneious v10.0.9 (*61*) restricting to reads with a mapping quality of >30. The Y haplogroup was determined by yHaplo program (*62*). For each male individual, we used a pileup of 13,508 International Society of Genetic Genealogy (ISOGG) SNPs (strand-ambiguous ones were excluded) and randomly drew a single base representing the genotype at each SNP position, with the same quality filtering applied to genotyping autosomes. For the genetic relatedness, we calculated pairwise mismatch rates of pseudodiploid genotypes across all SNPs. In addition, we applied the software READ (*63*), which confirmed the kinship estimates from the pairwise mismatch rate. This analysis revealed that the two petrous bones from samples LUK001 and LUK002 are from the same individual, and we merged the two libraries.

Present-day human data and published ancient genomes

We merged our newly reported ancient genomes published ancient African genomes (*2*–*4*, *9*–*11*), together with 584 individuals from 59 modern African populations (*1*, *12*, *64*, *65*) genotyped on the Affymetrix Human Origins array (Human Origins), and high-coverage genomes from the Simons Genome Diversity Project (*13*, *14*), including 300 individuals from 142 worldwide populations and 44 individuals from 22 African indigenous populations. Intersecting with SNPs present in the Human Origins array, we obtain data for 593,124 autosomal SNPs across worldwide populations.

PCA and admixture-clustering analyses

We used smartpca v16000 from the EIGENSOFT v7.2.1 package (*66*) for PCA using all autosome SNPs and projected ancient individuals on eigenvectors computed from present-day African populations on the Affymetric Human Origin array with option "lsqproject: YES" on. We used ADMIXTURE v1.3.0 (*18*) for unsupervised genetic clustering analysis of ancient African samples along with present-day Africans and all published ancient Africans and Levant Neolithic individuals. One individual, NYA003, was removed from ADMIXTURE analysis due to its second-degree relationship with NYA002.

Outgroup *f3* tests and symmetry *f4* tests

We performed outgroup *f3* with chimpanzee as outgroup, to check how our samples are closely related to present-day Africans and West Eurasians. The *f3* and *f4* statistics were calculated using the qp3Pop (v400) and qpDstat (v711) programs in the AdmixTools v5.1 package (*64*). We also performed model-based *f4* statistics for testing an additional genetic component in ancient eastern African Pastoralist groups (fig. S3). We used an in-house script that was first applied in (*10*) to compute *f4* statistics in form of (outgroup, test additional source group; two-way admixture model, Target). We used Ethiopia_4500BP + Levant_ChL as the hypothesized two-way admixture model and Dinka as the test additional source group and calculated model-based *f4* with varying Ethiopia_4500BP–related proportion in the two-way model from 0 to 100% in increments of 0.1%, with SE added by 5-centimorgan block jackknife method.

qpWave and qpAdm analyses

For modeling ancestral components, we used qpWave v410 and qpAdm v810 (*65*) in the AdmixTools v5.1 package (*64*). Here, we used transversions only to avoid the artifact from aDNA fragments, and with "allsnps: YES" option on to maximize the allele frequency–based resolution. We use a set of 12 worldwide populations—Mbuti, Mende, Dinka, Khomani, Anatolia_Neolithic, Iran_Ganj_Dareh_Neolithic_published, French, Sardinian, Punjabi, Ami, Onge, and Karitiana—as outgroup in our test and move a certain population from the outgroup list into the reference population list if needed.

Dating admixture

We used DATES v600 (*67*) for dating individual- and group-based admixture. A default bin size of 0.001 Morgans is applied in our estimates (flag "binsize: 0.001" added).

Phenotypic SNP analyses

We examined SNPs encoding for biological traits in the newly reported ancient African genomes, such as LP, Malaria resistance, and eye/skin pigmentation, following the list of SNPs used in (*68*). For each phenotype-associated locus, we report the number of reads with derived alleles versus the total number of reads covered on this site in table S7, by applying SAMtools pileup on BAM files after quality filtering (*-q 30 -Q 30*).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/24/eaaz0183/DC1

REFERENCES AND NOTES

1. J. K. Pickrell, N. Patterson, C. Barbieri, F. Berthold, L. Gerlach, T. Güldemann, B. Kure, S. W. Mpoloka, H. Nakagawa, C. Naumann, M. Lipson, P.-R. Loh, J. Lachance, J. Mountain, C. D. Bustamante, B. Berger, S. A. Tishkoff, B. M. Henn, M. Stoneking, D. Reich, B. Pakendorf, The genetic prehistory of southern Africa. *Nat. Commun.* **3**, 1143 (2012).
2. C. M. Schlebusch, H. Malmström, T. Günther, P. Sjödin, A. Coutinho, H. Edlund, A. R. Munters, M. Vicente, M. Steyn, H. Soodyall, M. Lombard, M. Jakobsson, Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* **358**, 652–655 (2017).
3. P. Skoglund, J. C. Thompson, M. E. Prendergast, A. Mittnik, K. Sirak, M. Hajdinjak, T. Salie, N. Rohland, S. Mallick, A. Peltzer, A. Heinze, I. Olalde, M. Ferry, E. Harney, M. Michel, K. Stewardson, J. I. Cerezo-Román, C. Chiumia, A. Crowther, E. Gomani-Chindebvu, A. O. Gidna, K. M. Grillo, I. T. Helenius, G. Hellenthal, R. Helm, M. Horton, S. López, A. Z. P. Mabulla, J. Parkington, C. Shipton, M. G. Thomas, R. Tibesasa, M. Welling, V. M. Hayes, D. J. Kennett, R. Ramesar, M. Meyer, S. Pääbo, N. Patterson, A. G. Morris, N. Boivin, R. Pinhasi, J. Krause, D. Reich, Reconstructing prehistoric African population structure. *Cell* **171**, 59–71.e21 (2017).
4. M. E. Prendergast, M. Lipson, E. A. Sawchuk, I. Olalde, C. A. Ogola, N. Rohland, K. A. Sirak, N. Adamski, R. Bernardos, N. Broomandkhoshbacht, K. Callan, B. J. Culleton, L. Eccles, T. K. Harper, A. M. Lawson, M. Mah, J. Oppenheimer, K. Stewardson, F. Zalzala, S. H. Ambrose, G. Ayodo, H. L. Gates Jr., A. O. Gidna, M. Katongo, A. Kwekason, A. Z. P. Mabulla, G. S. Mudenda, E. K. Ndiema, C. Nelson, P. Robertshaw, D. J. Kennett, F. K. Manthi, D. Reich, Ancient DNA reveals a multistep spread of the first herders into sub-Saharan Africa. *Science* **365**, eaaw6275 (2019).
5. R. Pinhasi, D. Fernandes, K. Sirak, M. Novak, S. Connell, S. Alpaslan-Roodenberg, F. Gerritsen, V. Moiseyev, A. Gromov, P. Raczky, A. Anders, M. Pietrusewsky, G. Rollefson, M. Jovanovic, H. Trinhhoang, G. Bar-Oz, M. Oxenham, H. Matsumura, M. Hofreiter, Optimal ancient DNA yields from the inner ear part of the human petrous bone. *PLOS One* **10**, e0129102 (2015).
6. M. G. Llorente, E. R. Jones, A. Eriksson, V. Siska, K. W. Arthur, J. W. Arthur, M. C. Curtis, J. T. Stock, M. Coltorti, P. Pieruccini, S. Stretton, F. Brock, T. Higham, Y. Park, M. Hofreiter, D. G. Bradley, J. Bhak, R. Pinhasi, A. Manica, Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science* **350**, 820–822 (2015).
7. T. Guldemann, A linguist's view: Khoe-Kwadi speakers as the earliest food-producers of southern Africa. *South. Afr. Humanit.* **20**, 93–132 (2008).
8. J. K. Pickrell, N. Patterson, P.-R. Loh, M. Lipson, B. Berger, M. Stoneking, B. Pakendorf, D. Reich, Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2632–2637 (2014).
9. M. G. Llorente, E. R. Jones, A. Eriksson, V. Siska, K. W. Arthur, J. W. Arthur, M. C. Curtis, J. T. Stock, M. Coltorti, P. Pieruccini, S. Stretton, F. Brock, T. Higham, Y. Park, M. Hofreiter, D. G. Bradley, J. Bhak, R. Pinhasi, A. Manica, Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. *Science* **350**, 820–822 (2015).
10. M. van de Loosdrecht, A. Bouzouggar, L. Humphrey, C. Posth, N. Barton, A. Aximu-Petri, B. Nickel, S. Nagel, E. H. Talbi, M. A. El Hajraoui, S. Amzazi, J.-J. Hublin, S. Pääbo, S. Schiffels, M. Meyer, W. Haak, C. Jeong, J. Krause, Pleistocene North African genomes link near Eastern and sub-Saharan African human populations. *Science* **360**, 548–552 (2018).
11. R. Fregel, F. L. Méndez, Y. Bokbot, D. Martín-Socas, M. D. Camalich-Massieu, J. Santana, J. Morales, M. C. Ávila-Arcos, P. A. Underhill, B. Shapiro, G. Wojcik, M. Rasmussen, A. E. R. Soares, J. Kapp, A. Sockell, F. J. Rodríguez-Santos, A. Mikdad, A. Trujillo-Mederos, C. D. Bustamante, Ancient genomes from north Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 6774–6779 (2018).
12. I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, B. Berger, C. Economou, R. Bollongino, Q. Fu, K. I. Bos, S. Nordenfelt, H. Li, C. de Filippo, K. Prüfer, S. Sawyer, C. Posth, W. Haak, F. Hallgren, E. Fornander, N. Rohland, D. Delsate, M. Francken, J.-M. Guinet, J. Wahl, G. Ayodo, H. A. Babiker, G. Bailliet, E. Balanovska, O. Balanovsky, R. Barrantes, G. Bedoya, H. Ben-Ami, J. Bene, F. Berrada, C. M. Bravi, F. Brisighelli, G. B. J. Busby, F. Cali, M. Churnosov, D. E. C. Cole, D. Corach, L. Damba, G. van Driem, S. Dryomov, J.-M. Dugoujon, S. A. Fedorova, I. G. Romero, M. Gubina, M. Hammer, B. M. Henn, T. Hervig, U. Hodoglugil, A. R. Jha, S. Karachanak-Yankova, R. Khusainova, E. Khusnutdinova, R. Kittles, T. Kivisild, W. Klitz, V. Kučinskas, A. Kushniarevich, L. Laredj, S. Litvinov, T. Loukidis, R. W. Mahley, B. Melegh, E. Metspalu, J. Molina, J. Mountain, K. Näkkäläjärvi, D. Nesheva, T. Nyambo, L. Osipova, J. Parik, F. Platonov, O. Posukh, V. Romano, F. Rothhammer, I. Rudan, R. Ruizbakiev, H. Sahakyan, A. Sajantila, A. Salas, E. B. Starikovskaya, A. Tarekegn, D. Toncheva, S. Turdikulova, I. Uktveryte, O. Utevska, R. Vasquez, M. Villena, M. Voevoda, C. A. Winkler, L. Yepiskoposyan, P. Zalloua, T. Zemunik, A. Cooper, C. Capelli, M. G. Thomas, A. Ruiz-Linares, S. A. Tishkoff, L. Singh, K. Thangaraj, R. Villems, D. Comas, R. Sukernik, M. Metspalu, M. Meyer, E. E. Eichler, J. Burger, M. Slatkin, S. Pääbo, J. Kelso, D. Reich, J. Krause, Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
13. S. Fan, D. E. Kelly, M. H. Beltrame, M. E. B. Hansen, S. Mallick, A. Ranciaro, J. Hirbo, S. Thompson, W. Beggs, T. Nyambo, S. A. Omar, D. W. Meskel, G. Belay, A. Froment, N. Patterson, D. Reich, S. A. Tishkoff, African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biol.* **20**, 82 (2019).
14. S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, D. Reich, The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
15. G. Renaud, V. Slon, A. T. Duggan, J. Kelso, Schmutzi: Estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* **16**, 224 (2015).
16. Q. Fu, A. Mittnik, P. L. F. Johnson, K. Bos, M. Lari, R. Bollongino, C. Sun, L. Giemsch, R. Schmitz, J. Burger, A. M. Ronchitelli, F. Martini, R. G. Cremonesi, J. Svoboda, P. Bauer, D. Caramelli, S. Castellano, D. Reich, S. Pääbo, J. Krause, A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* **23**, 553–559 (2013).
17. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
18. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
19. W. Haak, I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, K. Stewardson, Q. Fu, A. Mittnik, E. Bánffy, C. Economou, M. Francken, S. Friederich, R. G. Pena, F. Hallgren, V. Khartanovich, A. Khokhlov, M. Kunst, P. Kuznetsov, H. Meller, O. Mochalov, V. Moiseyev, N. Nicklisch, S. L. Pichler, R. Risch, M. A. R. Guerra, C. Roth, A. Szécsényi-Nagy, J. Wahl, M. Meyer, J. Krause, D. Brown, D. Anthony, A. Cooper, K. W. Alt, D. Reich, Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
20. B. M. Henn, L. R. Botigué, S. Gravel, W. Wang, A. Brisbin, J. K. Byrnes, K. Fadhlaoui-Zid, P. A. Zalloua, A. Moreno-Estrada, J. Bertranpetit, C. D. Bustamante, D. Comas, Genomic ancestry of North Africans supports back-to-Africa migrations. *PLOS Genet.* **8**, e1002397 (2012).
21. R. Kuper, S. Kröpelin, Climate-controlled holocene occupation in the sahara: Motor of Africa's evolution. *Science* **313**, 803–807 (2006).
22. A. G. Morris, The myth of the east african `Bushmen'. *South Afr. Archaeol. Bull.* **58**, 85–90 (2003).
23. T. Güldemann, Greenberg's" case" for Khoisan: The morphological evidence, in *Problems of Linguistic-Historical Reconstruction in Africa*, D. Ibriszimov, Ed. (Köln: Rüdiger Köppe, 2008), pp.123–153.
24. É. Harney, H. May, D. Shalem, N. Rohland, S. Mallick, I. Lazaridis, R. Sarig, K. Stewardson, S. Nordenfelt, N. Patterson, I. Hershkovitz, D. Reich, Ancient DNA from chalcolithic israel reveals the role of population mixture in cultural transformation. *Nat. Commun.* **9**, 3336 (2018).

25. S. H. Ambrose, Chronology of the later stone age and food production in east africa. *J. Archaeol. Sci.* **25**, 377–392 (1998).
26. P. J. Lane, The "Moving Frontier" and the transition to food production in Kenya. *Azania* **39**, 243–264 (2004).
27. D. Gifford-Gonzalez, Early pastoralists in east africa: Ecological and social dimensions. *J. Anthropol. Archaeol.* **17**, 166–200 (1998).
28. M. E. Prendergast, K. K. Mutundu, Late holocene archaeological faunas in east Africa: Ethnographic analogues and interpretive challenges. *Documenta Archaeobiologiae* **7**, 203–232 (2009).
29. J. C. Onyango-Abuje, Crescent island: A preliminary report on excavations at an east african neolithic site. *Azania Archaeol. Res. Africa* **12**, 147–159 (1977).
30. D. P. Gifford, G. L. Isaac, C. M. Nelson, Evidence for predation and pastoralism at prolonged drift: A pastoral neolithic site in kenya. *Azania* **15**, 57–108 (1980).
31. A. B. Smith, Keeping people on the periphery: The ideology of social hierarchies between hunters and herders. *J. Anthropol. Archaeol.* **17**, 201–215 (1998).
32. V. Bajić, C. Barbieri, A. Hübner, T. Güldemann, C. Naumann, L. Gerlach, F. Berthold, H. Nakagawa, S. W. Mpoloka, L. Roewer, J. Purps, M. Stoneking, B. Pakendorf, Genetic structure and sex-biased gene flow in the history of southern african populations. *Am. J. Phys. Anthropol.* **167**, 656–671 (2018).
33. B. A. Ogot, *History of the Southern Luo. Volume 1. Migration and Settlement, 1500-1900* (East African Publishing House, 1967).
34. P. Robertshaw, Munsa earthworks: A preliminary report on recent excavations. *Azani Arch. Res. Africa* **32**, 1–20 (1997).
35. S. A. Tishkoff, M. K. Gonder, B. M. Henn, H. Mortensen, A. Knight, C. Gignoux, N. Fernandopulle, G. Lema, T. B. Nyambo, U. Ramakrishnan, F. A. Reed, J. L. Mountain, History of click-speaking populations of africa inferred from mtDNA and Y chromosome genetic variation. *Mol. Biol. Evol.* **24**, 2180–2195 (2007).
36. C. M. Schlebusch, T. Naidoo, H. Soodyall, SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa. *Electrophoresis* **30**, 3657–3664 (2009).
37. B. M. Henn, C. Gignoux, A. A. Lin, P. J. Oefner, P. Shen, R. Scozzari, F. Cruciani, S. A. Tishkoff, J. L. Mountain, P. A. Underhill, Y-chromosomal evidence of a pastoralist migration through tanzania to southern Africa. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 10693–10698 (2008).
38. N. Isern, J. Fort, Assessing the importance of cultural diffusion in the Bantu spread into southeastern Africa. *PLOS One* **14**, e0215573 (2019).
39. G. Breton, C. M. Schlebusch, M. Lombard, P. Sjödin, H. Soodyall, M. Jakobsson, Lactase persistence alleles reveal partial east african ancestry of southern african Khoe pastoralists. *Curr. Biol.* **24**, 852–858 (2014).
40. E. Macholdt, V. Lede, C. Barbieri, S. W. Mpoloka, H. Chen, M. Slatkin, B. Pakendorf, M. Stoneking, Tracing pastoralist migrations to southern Africa with lactase persistence alleles. *Curr. Biol.* **24**, 875–879 (2014).
41. N. Sepúlveda, A. Manjurano, S. G. Campino, M. Lemnge, J. Lusingu, R. Olomi, K. A. Rockett, C. Hubbart, A. Jeffreys, K. Rowlands, T. G. Clark, E. M. Riley, C. J. Drakeley; MalariaGEN Consortium, Malaria host candidate genes validated by association with current, recent, and historical measures of transmission intensity. *J Infect Dis.* **216**, 45–54 (2017).
42. K. A. Murphy, A meal on the hoof or wealth in the kraal? Stable isotopes at Kgaswe and Taukome in eastern Botswana. *Int. J. Osteoarchaeol.* **21**, 591–601 (2011).
43. G. Turner, Early iron age herders in northwestern Botswana: The faunal evidence. *Botsw. Notes Rec.* **19**, 7–23 (1987).
44. J. K. Thornton, L. Heywood, Afro-Latino Voices, *Narratives from the Early Modern Ibero-Atlantic World, 1550-1812*, K. J. McKnight, L. J. Garofalo, Eds. (Hackett Publishing, 2009).
45. J. E. Yellen, Barbed bone points: Tradition and continuity in Saharan and sub-Saharan Africa. *African Arch. Rev.* **15**, 173–198 (1998).
46. B. Keding, Middle holocene fisher-hunter-gatherers of lake turkana in Kenya and their cultural connections with the north: The pottery. *J. African Arch.* **15**, 42–76 (2017).
47. A. Crowther, M. E. Prendergast, D. Q. Fuller, N. Boivin, Subsistence mosaics, forager-farmer interactions, and the transition to food production in eastern Africa. *Quat. Int.* **489**, 101–120 (2018).
48. P. Mitchell, Early farming communities of southern and south-central Africa, in *The Oxford Handbook of African Archaeology*, P. Mitchell, P. Lane, Eds. (Oxford Univ. Press, 2013), pp. 657–670.
49. S. A. Tishkoff, F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A. A. Awomoyi, J.-M. Bodo, O. Doumbo, M. Ibrahim, A. T. Juma, M. J. Kotze, G. Lema, J. H. Moore, H. Mortensen, T. B. Nyambo, S. A. Omar, K. Powell, G. S. Pretorius, M. W. Smith, M. A. Thera, C. Wambebe, J. L. Weber, S. M. Williams, The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044 (2009).
50. P. J. Reimer, E. Bard, A. Bayliss, J. Warren Beck, P. G. Blackwell, C. B. Ramsey, C. E. Buck, H. Cheng, R. Lawrence Edwards, M. Friedrich, P. M. Grootes, T. P. Guilderson, H. Haflidason, I. Hajdas, C. Hatté, T. J. Heaton, D. L. Hoffmann, A. G. Hogg, K. A. Hughen, K. Felix Kaiser, B. Kromer, S. W. Manning, M. Niu, R. W. Reimer, D. A. Richards, E. Marian Scott, J. R. Southon, R. A. Staff, C. S. M. Turney, J. van der Plicht, IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon* **55**, 1869–1887 (2013).
51. C. Bronk Ramsey, T. F. G. Higham, F. Brock, D. Baker, P. Ditchfield, Radiocarbon dates from the Oxford AMS system: *Archaeometry* Datelist 33. *Archaeometry* **51**, 323–349 (2009).
52. J. Dabney, M. Knapp, I. Glocke, M.-T. Gansauge, A. Weihmann, B. Nickel, C. Valdiosera, N. García, S. Pääbo, J.-L. Arsuaga, M. Meyer, Complete mitochondrial genome sequence of a middle pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15758–15763 (2013).
53. N. Rohland, E. Harney, S. Mallick, S. Nordenfelt, D. Reich, Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20130624 (2015).
54. Q. Fu, M. Hajdinjak, O. T. Moldovan, S. Constantin, S. Mallick, P. Skoglund, N. Patterson, N. Rohland, I. Lazaridis, B. Nickel, B. Viola, K. Prüfer, M. Meyer, J. Kelso, D. Reich, S. Pääbo, An early modern human from Romania with a recent neanderthal ancestor. *Nature* **524**, 216–219 (2015).
55. A. Peltzer, G. Jäger, A. Herbig, A. Seitz, C. Kniep, J. Krause, K. Nieselt, EAGER: Efficient ancient genome reconstruction. *Genome Biol.* **17**, 60 (2016).
56. M. Schubert, S. Lindgreen, L. Orlando, AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC. Res. Notes* **9**, 88 (2016).
57. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
58. G. Jun, M. K. Wing, G. R. Abecasis, H. M. Kang, An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.* **25**, 918–925 (2015).
59. H. Weissensteiner, D. Pacher, A. Kloss-Brandstätter, L. Forer, G. Specht, H.-J. Bandelt, F. Kronenberg, A. Salas, S. Schönherr, HaploGrep 2: Mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58–W63 (2016).
60. D. Vianello, F. Sevini, G. Castellani, L. Lomartire, M. Capri, C. Franceschi, HAPLOFIND: A new method for high-throughput mtDNA haplogroup assignment. *Hum. Mutat.* **34**, 1189–1194 (2013).
61. M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, A. Drummond, Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
62. G. David Poznik, Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. *bioRxiv*, 088716 (2016).
63. J. M. Monroy Kuhn, M. Jakobsson, T. Günther, Estimating genetic kin relationships in prehistoric populations. *PLOS One* **13**, e0195491 (2018).
64. N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, D. Reich, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
65. I. Lazaridis, D. Nadel, G. Rollefson, D. C. Merrett, N. Rohland, S. Mallick, D. Fernandes, M. Novak, B. Gamarra, K. Sirak, S. Connell, K. Stewardson, E. Harney, Q. Fu, G. Gonzalez-Fortes, E. R. Jones, S. A. Roodenberg, G. Lengyel, F. Bocquentin, B. Gasparian, J. M. Monge, M. Gregg, V. Eshed, A.-S. Mizrahi, C. Meiklejohn, F. Gerritsen, L. Bejenaru, M. Blüher, A. Campbell, G. Cavalleri, D. Comas, P. Froguel, E. Gilbert, S. M. Kerr, P. Kovacs, J. Krause, D. McGettigan, M. Merrigan, D. A. Merriwether, S. O'Reilly, M. B. Richards, O. Semino, M. Shamoon-Pour, G. Stefanescu, M. Stumvoll, A. Tönjes, A. Torroni, J. F. Wilson, L. Yengo, N. A. Hovhannisyan, N. Patterson, R. Pinhasi, D. Reich, Genomic insights into the origin of farming in the ancient near east. *Nature* **536**, 419–424 (2016).
66. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLOS Genet.* **2**, e190 (2006).
67. V. M. Narasimhan, N. Patterson, P. Moorjani, I. Lazaridis, M. Lipson, S. Mallick, N. Rohland, R. Bernardos, A. M. Kim, N. Nakatsuka, I. Olalde, A. Coppa, J. Mallory, V. Moiseyev, J. Monge, L. M. Olivieri, N. Adamski, N. Broomandkhoshbacht, F. Candilio, O. Cheronet, B. J. Culleton, M. Ferry, D. Fernandes, B. Gamarra, D. Gaudio, M. Hajdinjak, É. Harney, T. K. Harper, D. Keating, A. M. Lawson, M. Michel, M. Novak, J. Oppenheimer, N. Rai, K. Sirak, V. Slon, K. Stewardson, Z. Zhang, G. Akhatov, A. N. Bagashev, B. Baitanayev, G. L. Bonora, T. Chikisheva, A. Derevianko, E. Dmitry, K. Douka, N. Dubova, A. Epimakhov, S. Freilich, D. Fuller, A. Goryachev, A. Gromov, B. Hanks, M. Judd, E. Kazizov, A. Khokhlov, E. Kitov, E. Kupriyanova, P. Kuznetsov, D. Luiselli, F. Maksudov, C. Meiklejohn, D. Merrett, R. Micheli, O. Mochalov, Z. Muhammed, S. Mustafokulov, A. Nayak, R. M. Petrovna, D. Pettener, R. Potts, D. Razhev, S. Sarno, K. Sikhymbaeva, S. M. Slepchenko, N. Stepanova, S. Svyatko, S. Vasilyev, M. Vidale, D. Voyakin, A. Yermolayeva, A. Zubova, V. S. Shinde, C. Lalueza-Fox, M. Meyer, D. Anthony, N. Boivin, K. Thangaraj, D. J. Kennett, M. Frachetti, R. Pinhasi, D. Reich, The genomic formation of South and Central Asia. bioRxiv 292581 [**Preprint**] (31 March 2018).
68. M. Feldman, D. M. Master, R. A. Bianco, M. Burri, P. W. Stockhammer, A. Mittnik, A. J. Aja, C. Jeong, J. Krause, Ancient DNA sheds light on the genetic origins of early iron age philistines. *Sci. Adv.* **5**, eaax0061 (2019).

69. C. A. Tryon, I. Crevecoeur, J. T. Faith, R. Ekshtain, J. Nivens, D. Patterson, E. N. Mbua, F. Spoor, Late pleistocene age and archaeological context for the hominin calvaria from GvJm-22 (Lukenya Hill, Kenya). *Proc. Natl. Acad. Sci. U.S.A.* **112**, 2682–2687 (2015).
70. F. Marshall, R. E. B. Reid, S. Goldstein, M. Storozum, A. Wreschnig, L. Hu, P. Kiura, R. Shahack-Gross, S. H. Ambrose, Ancient herders enriched and restructured African grasslands. *Nature* **561**, 387–390 (2018).
71. C. M. Nelson, J. Kimegich, in *Origin and Early Development of Food – Producing Cultures in North-Eastern Africa* (Poznan Archaeological Museum, 1984) pp. 481–487.
72. S. H. Ambrose, M. J. DeNiro, Reconstruction of African human diet using bone collagen carbon and nitrogen isotope ratios. *Nature* **319**, 321–324 (1986).
73. E. A. Sawchuk, thesis, University of Toronto (2017).
74. L. A. Schepartz, thesis, University of Michigan (1987).
75. M. D. Leakey, L. S. B. Leakey, P. M. Game, A. J. H. Goodwin, Report on the excavations at Hyrax Hill, Nakuru, Kenya Colony, 1937–1938. *Trans. R. Soc. S. Afr.* **30**, 271–409 (1943).
76. E. A. Hildebrand, K. M. Grillo, E. A. Sawchuk, S. K. Pfeiffer, L. B. Conyers, S. T. Goldstein, A. C. Hill, A. Janzen, C. E. Klehm, M. Helper, P. Kiura, E. Ndiema, C. Ngugi, J. J. Shea, H. Wang, A monumental cemetery built by eastern Africa's first herders near Lake Turkana, Kenya. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 8942–8947 (2018).
77. H. Field, The University of California African expedition: II, Sudan and Kenya. *Am. Anthropol.* **51**, 72–84 (1949).
78. W. E. Owen, 76. The Early Smithfield culture of Kavirondo (Kenya) and South Africa. *Man.* **41**, 115 (1941).
79. J. L. Buckberry, A. T. Chamberlain, Age estimation from the auricular surface of the ilium: A revised method. *Am. J. Phys. Anthropol.* **119**, 231–239 (2002).
80. E. A. DiGangi, J. D. Bethard, E. H. Kimmerle, L. W. Konigsberg, A new method for estimating age-at-death from the first rib. *Am. J. Phys. Anthropol.* **138**, 164–176 (2009).
81. M. Trotter, R. R. Peterson, Weight of the skeleton during postnatal development. *Am. J. Phys. Anthropol.* **33**, 313–323 (1970).
82. E. C. Lanning, Ancient earthworks in western Uganda. *Uganda J.* **17**, 51–62 (1953).
83. P. Robertshaw, The age and function of ancient earthworks of western Uganda. *Uganda J.* **47**, 20–33 (2001).
84. E. C. Lanning, The munsa earthworks. *Uganda J.* **19**, 177–182 (1955).
85. L. Iles, P. Robertshaw, R. Young, A furnace and associated ironworking remains at Munsa, Uganda. *Azania Arch. Res. Africa* **49**, 45–63 (2014).
86. R. L. Tantala, thesis, University of Wisconsin, Madison (1989).
87. P. Robertshaw, The ancient earthworks of western Uganda: Capital sites of a Cwezi empire? *Uganda J.* **48**, 17–32 (2002).
88. J. Mercader, M. D. Garralda, O. M. Pearson, R. C. Bailey, Eight hundred-year-old human remains from the Ituri tropical forest, democratic republic of congo: The rock shelter site of Matangai Turu northwest. *Am. J. Phys. Anthropol.* **115**, 24–37 (2001).
89. J. Mercader, F. Runge, L. Vrydaghs, H. Doutrelepont, C. E. N. Ewango, J. Juan-Tresseras, Phytoliths from archaeological sites in the tropical forest of Ituri, democratic republic of congo. *Quatern. Res.* **54**, 102–112 (2000).
90. J. Mercader, S. Rovira, P. Gómez-Ramos, Forager-farmer interaction and ancient iron metallurgy in the Ituri rainforest, democratic republic of congo. *Azania Arch. Res. Africa* **35**, 107–122 (2000).
91. B. Clist, E. Cranshof, G.-M. de Schryver, D. Herremans, K. Karklins, I. Matonda, C. Polet, A. Sengeløv, F. Steyaert, C. Verhaeghe, K. Bostoen, The elusive archaeology of kongo urbanism: The case of kindoki, Mbanza Nsundi (Lower Congo, DRC). *African Arch. Rev.* **32**, 369–412 (2015).
92. B. Clist, E. Cranshof, P. de Maret, M. Kaumba Mazanga, R. Kidebua, I. Matonda, A. Nkanza Lutayi, J. Yogolelo, in *Une Archéologie des Provinces Septentrionales du Royaume Kongo* (Archaeopress, 2018), pp. 135–164.
93. B. Clist, N. Nikis, P. de Maret, in *Une Archéologie des Provinces Septentrionales du Royaume Kongo* (Archaeopress, 2018), pp. 243–295.
94. J. K. Thornton, in *The Kongo Kingdom: The Origins, Dynamics and Cosmopolitan Culture of an African Polity* (Cambridge Univ. Press, 2018), pp. 17–41.
95. C. Polet, in *Une archéologie des Provinces Septentrionales du Royaume Kongo* (Archeopress, 2018), pp. 401–438.
96. C. Polet, B.-O. Clist, K. Bostoen, Étude des restes humains de Kindoki (République démocratique du Congo, fin XVIIe–Début XIXe siècle). *Bull. Mém. Soc. Anthropol. Paris* **30**, 70–89 (2018).
97. C. Verhaeghe, B.-O. Clist, C. Fontaine, K. Karklins, K. Bostoen, W. De Clercq, Shell and glass beads from the tombs of Kindoki, Mbanza Nsundi, lower congo. *Beads J. Soc. Bead Res.* **26**, 23–34 (2014).
98. P. Dubrunfaut, B. Clist, in *Une Archéologie des Provinces Septentrionales du Royaume Kongo* (Archaeopress, 2018), pp. 359–368.
99. K. Karklins, B. Clist, in *Une Archéologie des Provinces Septentrionales du Royaume Kongo* (Archaeopress, 2018), pp. 337–348.
100. B. Clist, E. Cranshof, G.-M. de Schryver, D. Herremans, K. Karklins, I. Matonda, F. Steyaert, K. Bostoen, African-European contacts in the Kongo Kingdom (Sixteenth-eighteenth centuries): New archaeological insights from Ngongo Mbata (Lower Congo, DRC). *Int. J. Hist. Archaeol.* **19**, 464–501 (2015).
101. B. Clist, E. Cranshof, M. Kaumba Mazanga, I. Matonda Sakala, A. Nkanza Lutayi, J. Yogolelo, in *Une Archéologie des Provinces Septentrionales du Royaume Kongo* (Archaeopress, 2018), pp. 71–132.
102. M. Bequaert, Fouille d'un cimetière du XVIIe siècle au Congo Belge. *L'Antiquité Classique* **9**, 127–128 (1940).
103. E. Kose, New light on ironworking groups along the middle Kavango in northern Namibia. *South African Arch. Bull.* **64**, 130–147 (2009).
104. M. N. Mosothwane, Dietary stable carbon isotope signatures of the early iron age inhabitants of Ngamiland. *Botsw. Notes Rec.* **43**, 115–129 (2011).
105. E. N. Wilmsen, A. C. Campbell, G. A. Brook, L. H. Robbins, M. Murphy, Mining and moving specular haematite in Botswana, ca. 200–1300 AD, in *The World of Iron* (Archetype, 2013), pp. 33–45.
106. E. N. Wilmsen, Nqoma: An abridged review. *Botsw. Notes Rec.* **43**, 95–114 (2011).
107. J. R. Denbow, E. N. Wilmsen, Iron age pastoralist settlements in Botswana. *S. Afr. J. Sci.* **79**, 405–407 (1983).
108. J. Denbow, thesis, Indiana University (1983).
109. T. N. Huffman, *Handbook to the Iron Age* (University of KwaZulu-Natal Press, 2007).

Acknowledgments: We thank all the local collaborators and communities who were essential in the recovery of the recently excavated samples reported here. From Kakapel excavations, we are indebted to the communities of Kakoli, Aboloi, and Chelelemuk and the Busia County Commissioners Office. This project would not have been possible without the assistance of the staff and curators of the Nairobi National Museum and National Museums of Kenya. All research in Kenya was carried out under permits and permissions from the National Commission for Science, Technology, and Innovation, Kenya. We thank C. Polet for sampling help and T. Erler, R. Radzeviciute, A. Wissgott, and G. Brandt for help with sample preparation and aDNA laboratory work. We are grateful to the Efe and Lese community from Ngodingodi in the Ituri rainforest (DRC), without which we could not have completed this work. We thank G. Whitelaw for useful discussion about the arrival of Bantu speakers in South Africa. For the Walalde sampling, we thank the students who took part in the project for training and for their Master theses and also the Middle Senegal Valley population for hospitality. We thank J. Reinold as director of fieldwork conducted at Kadruka 1 and Kadruka 21. We also thanks M. Besse and J. Desideri of the Laboratoire d'archéologie préhistorique et anthropologie, University of Geneva for facilitating the sampling of the Kadruka material. **Funding:** The Botswana materials were excavated with the aid of a series (1979–1991) of NSF (USA) grants to E.N.W. The Walalde materials were collected during a research project funded by the NSF. S.S. and N.B. acknowledge funding from the Max Planck Society. The material from western DRC was excavated with funding from ERC-SG no. 284126 (2012–2016) and integrated here with funding from ERC-CG no. 724275. P.Robertsh. acknowledges funding of excavations in Uganda through an NSF (USA) grant. J.M.'s contribution was supported by the Canadian Social Sciences and Humanities Research Council through a Partnership Grant (serial no. 895-2016-1017). Part of the material from Senegal was procured within the Middle Senegal Valley Project (a joint project between Yale University and Université Cheikh Anta Diop) that was funded by the U.S. NSF grant 1534094. **Author contributions:** N.B. and S.S. conceived the study. S.G., E.S., M.B., A.C., R.C.P., E.N.W., B.C., A.D., K.B., J.M., C.O., E.N., P. Roberts, L.T.B., and A.D. collected and assembled skeletal material. S.G., M.B., E.S., A.C., E.N.W., B.C., K.B., A.D., J.M., L.T.B., R.C.P., R.J.M., C.O., E.N., P. Roberts, M.P., P.Robertsh., and N.B. provided archaeological and historical context. J.K. and S.S. supervised laboratory work and sequencing. K.W. and S.S. analyzed genetic data and, together with S.G., M.P., A.C., P. Roberts, and N.B., interpreted it in the context of archaeological information. K.W., S.G., N.B., and S.S. wrote the paper with input from all co-authors. **Competing interests:** The authors declare that they have no competing interests. **Date and materials availability:** The aligned sequences will be available via the European Nucleotide Archive under accession number PRJEB36063. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 6 August 2019
Accepted 15 April 2020
Published 12 June 2020
10.1126/sciadv.aaz0183

Citation: K. Wang, S. Goldstein, M. Bleasdale, B. Clist, K. Bostoen, P. Bakwa-Lufu, L. T. Buck, A. Crowther, A. Dème, R. J. McIntosh, J. Mercader, C. Ogola, R. C. Power, E. Sawchuk, P. Robertshaw, E. N. Wilmsen, M. Petraglia, E. Ndiema, F. K. Manthi, J. Krause, P. Roberts, N. Boivin, S. Schiffels, Ancient genomes reveal complex patterns of population movement, interaction, and replacement in sub-Saharan Africa. *Sci. Adv.* **6**, eaaz0183 (2020).

6. Manuscript C

CellPress
OPEN ACCESS

Cell

Article

A Dynamic 6,000-Year Genetic History of Eurasia's Eastern Steppe

Choongwon Jeong,[1,2,23,*] Ke Wang,[1,23] Shevan Wilkin,[3] William Timothy Treal Taylor,[3,4] Bryan K. Miller,[3,5] Jan H. Bemmann,[6] Raphaela Stahl,[1] Chelsea Chiovelli,[1] Florian Knolle,[1] Sodnom Ulziibayar,[7] Dorjpurev Khatanbaatar,[8] Diimaajav Erdenebaatar,[9] Ulambayar Erdenebat,[10] Ayudai Ochir,[11] Ganbold Ankhsanaa,[12] Chuluunkhuu Vanchigdash,[8] Battuga Ochir,[13] Chuluunbat Munkhbayar,[14] Dashzeveg Tumen,[10] Alexey Kovalev,[15] Nikolay Kradin,[16,17] Bilikto A. Bazarov,[17] Denis A. Miyagashev,[17] Prokopiy B. Konovalov,[17] Elena Zhambaltarova,[18] Alicia Ventresca Miller,[3,19] Wolfgang Haak,[1] Stephan Schiffels,[1] Johannes Krause,[1,20] Nicole Boivin,[3] Myagmar Erdene,[10] Jessica Hendy,[1,21] and Christina Warinner[1,20,22,24,*]

[1]Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena 07745, Germany
[2]School of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea
[3]Department of Archaeology, Max Planck Institute for the Science of Human History, Jena 07745, Germany
[4]Department of Anthropology, University of Colorado Boulder, Boulder, CO 80309, USA
[5]Museum of Anthropological Archaeology, University of Michigan, Ann Arbor, MI 48109, USA
[6]Department of Archaeology and Anthropology, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53113, Germany
[7]Institute of Archaeology, Mongolian Academy of Sciences, Ulaanbaatar 14200, Mongolia
[8]Mongolian University of Science and Technology, Ulaanbaatar 14191, Mongolia
[9]Department of Archaeology, Ulaanbaatar State University, Bayanzurkh district, Ulaanbaatar 13343, Mongolia
[10]Department of Anthropology and Archaeology, National University of Mongolia, Ulaanbaatar 14201, Mongolia
[11]International Institute for the Study of Nomadic Civilizations, Ulaanbaatar 14200, Mongolia
[12]National Centre for Cultural Heritage of Mongolia, Ulaanbaatar 14200, Mongolia
[13]Institute of History and Ethnology, Mongolian Academy of Sciences, Ulaanbaatar 14200, Mongolia
[14]University of Khovd, Khovd province, Khovd 84179, Mongolia
[15]Institute of Archaeology, Russian Academy of Sciences, Moscow 119991, Russia
[16]Institute of History, Archaeology and Ethnology, Far East Branch of the Russian Academy of Sciences, Vladivostok 690001, Russia
[17]Institute for Mongolian, Buddhist and Tibetan Studies, Siberian Branch of the Russian Academy of Sciences, Ulan-Ude 670047, Russia
[18]Department of Museology and Heritage, Faculty of Social and Cultural Activities, Heritage, and Tourism, Federal State Budgetary Educational Institution of Higher Education, East Siberian State Institute of Culture, Ulan-Ude 670031, Russia
[19]Department of Anthropology, University of Michigan, Ann Arbor, MI 48109, USA
[20]Faculty of Biological Sciences, Friedrich Schiller University, Jena 02134, Germany
[21]BioArCh, Department of Archaeology, University of York, York YO10 5NG, UK
[22]Department of Anthropology, Harvard University, Cambridge, MA 02138, USA
[23]These authors contributed equally
[24]Lead Contact
*Correspondence: cwjeong@snu.ac.kr (C.J.), warinner@fas.harvard.edu (C.W.)
https://doi.org/10.1016/j.cell.2020.10.015

SUMMARY

The Eastern Eurasian Steppe was home to historic empires of nomadic pastoralists, including the Xiongnu and the Mongols. However, little is known about the region's population history. Here, we reveal its dynamic genetic history by analyzing new genome-wide data for 214 ancient individuals spanning 6,000 years. We identify a pastoralist expansion into Mongolia ca. 3000 BCE, and by the Late Bronze Age, Mongolian populations were biogeographically structured into three distinct groups, all practicing dairy pastoralism regardless of ancestry. The Xiongnu emerged from the mixing of these populations and those from surrounding regions. By comparison, the Mongols exhibit much higher eastern Eurasian ancestry, resembling present-day Mongolic-speaking populations. Our results illuminate the complex interplay between genetic, sociopolitical, and cultural changes on the Eastern Steppe.

INTRODUCTION

Recent paleogenomic studies have revealed a dynamic population history on the Eurasian Steppe, with continental-scale migration events on the Western Steppe coinciding with Bronze Age transformations of Europe, the Near East, and the Caucasus (Allentoft et al., 2015; Damgaard et al., 2018a; 2018b; Haak et al., 2015; Mathieson et al., 2015; Wang et al., 2019). However, despite advances in understanding the genetic prehistory of the Western Steppe, the prehistoric population dynamics on

the Eastern Steppe remain poorly understood (Damgaard et al., 2018a; Jeong et al., 2018; Rogers, 2016). The Eastern Steppe is a great expanse of grasslands, forest steppe, and desert steppe extending more than 2,500 km (Figure 1; Figure S1). While also covering parts of modern-day China and Russia, most of the Eastern Steppe falls within the national boundaries of present-day Mongolia. Recent paleogenomic studies suggest that the eastern Eurasian forest steppe zone was genetically structured during the Pre-Bronze and Early Bronze Age periods, with a strong west-east admixture cline of ancestry stretching from Botai in central Kazakhstan to Lake Baikal in southern Siberia to Devil's Gate Cave in the Russian Far East (Damgaard et al., 2018a; Jeong et al., 2018; Sikora et al., 2019; Siska et al., 2017).

During the Bronze Age, the multi-phased introduction of pastoralism drastically changed lifeways and subsistence on the Eastern Steppe (Honeychurch, 2015; Kindstedt and Ser-Od, 2019). A recent large-scale paleoproteomic study has confirmed milk consumption in Mongolia prior to 2500 BCE by individuals affiliated with the Afanasievo (ca. 3000 BCE) and Chemurchek (2750–1900 BCE) cultures (Wilkin et al., 2020a). Although Afanasievo groups in the Upper Yenisei region have been genetically linked to the Yamnaya culture of the Pontic-Caspian steppe (ca. 3300–2200 BCE) (Allentoft et al., 2015; Morgunova and Khokhlova, 2013; Narasimhan et al., 2019), the origins of the Chemurchek have been controversial (Kovalev, 2014). Once introduced, ruminant dairying became widespread by the Middle/Late Bronze Age (MLBA, here defined as 1900–900 BCE), being practiced in the west and north at sites associated with Deer Stone-Khirigsuur Complex (DSKC) and in the east in association with the Ulaanzuukh culture (Jeong et al., 2018; Wilkin et al., 2020a). The relationships between DSKC and Ulaanzuukh groups are poorly understood, and little is known about other MLBA burial traditions in Mongolia, such as the Mönkhkhairkhan and Baitag. By the mid-first millennium BCE, the previous MLBA cultures were in decline, and Early Iron Age cultures emerged: the Slab Grave culture (ca. 1000–300 BCE) of eastern/southern Mongolia, whose burials sometimes incorporate uprooted materials from DSKC monuments (Fitzhugh, 2009; Honeychurch, 2015; Tsybiktarov, 2003; Volkov, 2002), and the Sagly/Uyuk culture (ca. 500–200 BCE) of the Sayan mountains to the northwest (also known as the Sagly-Bazhy culture, or Chandman culture in Mongolia), who had strong cultural ties to the Pazyryk (ca. 500–200 BCE) and Saka (ca. 900–200 BCE) cultures of the Altai and eastern Kazakhstan (Savinov, 2002; Tseveendorj, 2007).

From the late first millennium BCE onward, a series of hierarchical and centrally organized empires arose on the Eastern Steppe, notably the Xiongnu (209 BCE–98 CE), Türkic (552–742 CE), Uyghur (744–840 CE), and Khitan (916–1125 CE) empires. The Xiongnu empire was the first such polity in the steppe, whose drastic expansions into northern China, southern Siberia, and deep into Central Asia had a profound impact on the demographics and geopolitics of Eurasia. The Mongol empire, emerging in the thirteenth century CE, was the last and most expansive of these regimes, eventually controlling vast territories and trade routes stretching from China to the Mediterranean. However, due to a lack of large-scale genetic studies, the origins and relationships of the people who formed these states, including both the ruling elites and local commoners, remain obscure.

To clarify the population dynamics on the Eastern Steppe since prehistory, we generated and analyzed genome-wide genetic datasets for 214 individuals from 85 Mongolian and 3 Russian sites spanning approximately 6,000 years of time (ca. 4600 BCE to 1400 CE) (Tables S1, S2, and S3A). To this, we added recently published genomic data for 19 Bronze Age individuals from northern Mongolia (Jeong et al., 2018), as well as datasets from neighboring ancient populations in Russia and Kazakhstan (Damgaard et al., 2018a; 2018b; Narasimhan et al., 2019; Sikora et al., 2019; Unterländer et al., 2017) (Tables S3B and S3C), which we analyze together with worldwide modern reference populations (Table S3C). We also generated 30 new accelerator mass spectrometry dates, supplementing 74 previously published radiocarbon dates (Jeong et al., 2018; Taylor et al., 2019), for a total of 98 directly dated individuals (104 total dates) in this study (Table S4).

RESULTS

Pre-Bronze Age Population Structure and the Arrival of Pastoralism

In this study, we analyzed six pre-Bronze Age hunter-gatherer individuals from three sites dating to the fifth and fourth millennia BCE: one from eastern Mongolia (SOU001, "eastMongolia_preBA," 4686–4495 cal. BCE), one from central Mongolia (ERM003, "centralMongolia_preBA," 3781–3639 cal. BCE), and four from the eastern Baikal region ("Fofonovo_EN"). By comparing these genomes to previously published ancient and modern data across Eurasia (Figure 2; Table S3C), we found that they are most closely related to contemporaneous hunter-gatherers from the western Baikal region ("Baikal_EN," 5200–4200 BCE) and the Russian Far East ("DevilsCave_N," ca. 5700 BCE), filling in the geographic gap in the distribution of this genetic profile (Figure 3A). We refer to this profile as "Ancient Northeast Asian" (ANA) to reflect its geographic distribution relative to another widespread mid-Holocene genetic profile known as "Ancient North Eurasian" (ANE), which is found among the Pleistocene hunter-gatherers of the Mal'ta (ca. 24500–24100 BP) and Afontova Gora (ca. 16900–16500 BP) sites in Siberia (Fu et al., 2016; Raghavan et al., 2015) and the horse-herders of Botai, Kazakhstan (ca. 3500–3300 BCE) (Damgaard et al., 2018a). In principal component analysis (PCA) (Figure 2), ancient ANA individuals fall close to the cluster of present-day Tungusic- and Nivkh-speaking populations in northeast Asia, indicating that their genetic profile is still present in indigenous populations of the Far East today (Figure S3A). EastMongolia_preBA is genetically indistinguishable from the ANA group DevilsCave_N (Figures 3A and 4A; Figure S4A; Table S5A), whereas Fofonovo_EN and the slightly later centralMongolia_preBA both derive a minority (12%–17%) of their ancestry from ANE-related (Botai-like) groups with the remainder of their ancestry (83%–87%) characterized as ANA (Figures 3A and 4A; Table S5A). Re-analyzing published data from the western Baikal early Neolithic Kitoi culture (Baikal_EN) and the early Bronze Age Glazkovo culture (Baikal_EBA) (Damgaard et al., 2018a), we find that they

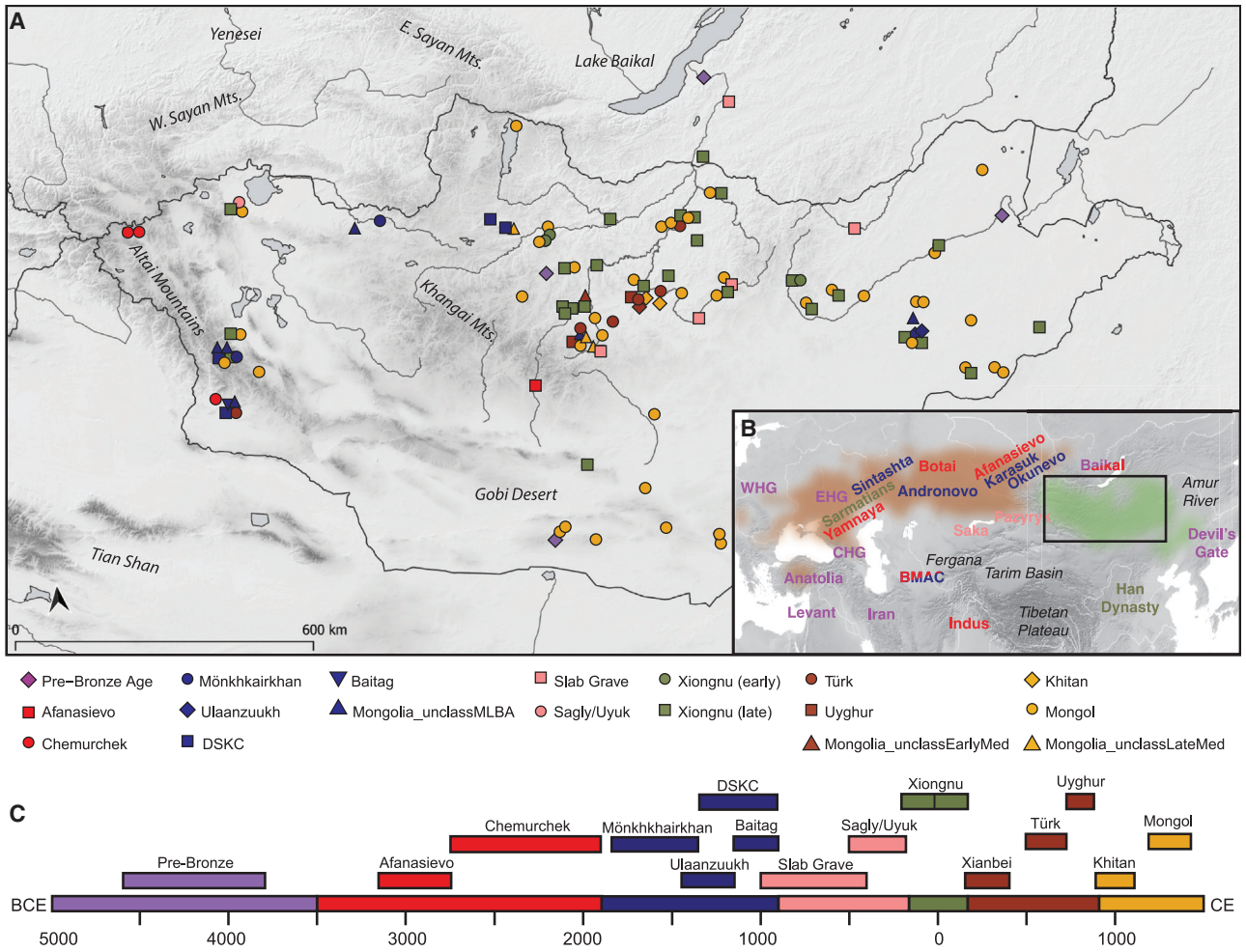


Figure 1. Overview of Ancient Populations and Time Periods

(A) Distribution of sites with their associated culture and time period indicated by color: Pre-Bronze, purple; Early Bronze, red; Middle/Late Bronze, blue; Early Iron, pink; Xiongnu, green; Early Medieval, brown; Late Medieval, gold (see STAR Methods). See Figure S1A and Table S1B for site codes and labels.

(B) Inset map of Eurasia indicating area of present study (box) and the locations of other ancient populations referenced in the text, colored by time period. The geographic extent of the Western/Central Steppe is indicated in light brown, and the Eastern Steppe is indicated in light green.

(C) Timeline of major temporal periods and archaeological cultures in Mongolia. Site locations have been jittered to improve visibility of overlapping sites.

have similar ancestry profiles and a slight increase in ANE ancestry through time (from 6.4% to 20.1%) (Figure 3A).

Pastoralism in Mongolia is often assumed to have been introduced by the eastward expansion of Western Steppe cultures (e.g., Afanasievo) via either the Upper Yenisei and Sayan mountain region to the northwest of Mongolia or through the Altai mountains in the west (Janz et al., 2017). Although the majority of Afanasievo burials reported to date are located in the Altai mountains and Upper Yenisei regions, the Early Bronze Age (EBA) site of Shatar Chuluu in the southern Khangai Mountains of central Mongolia has yielded Afanasievo-style graves with proteomic evidence of ruminant milk consumption (Wilkin et al., 2020a) and a western Eurasian mitochondrial haplogroup (Rogers et al., 2020). Analyzing two of these individuals (Afanasievo_Mongolia, 3112–2917 cal. BCE), we find that their genetic profiles are indistinguishable from that of published Afanasievo individuals from the Yenisei region (Allentoft et al., 2015; Narasimhan et al., 2019) (Figure 2; Figure S5C; Table S5B), and thus these two Afanasievo individuals confirm that the EBA expansion of Western Steppe herders (WSH) extended a further 1,500 km eastward beyond the Altai into the heart of central Mongolia (Figure 3A).

The succeeding EBA Chemurchek culture (2750–1900 BCE), a ruminant dairying society (Wilkin et al., 2020a) whose mortuary features include stone slabs and anthropomorphic stelae, has also been purportedly linked to WSH migrations (Kovalev and Erdenebaatar, 2009). Chemurchek graves are found throughout the Altai and in the Dzungarian Basin in Xinjiang, China (Jia and Betts, 2010; Kovalev, 2014; 2015). We analyzed two Chemurchek individuals from the southern Altai site of Yagshiin Huduu

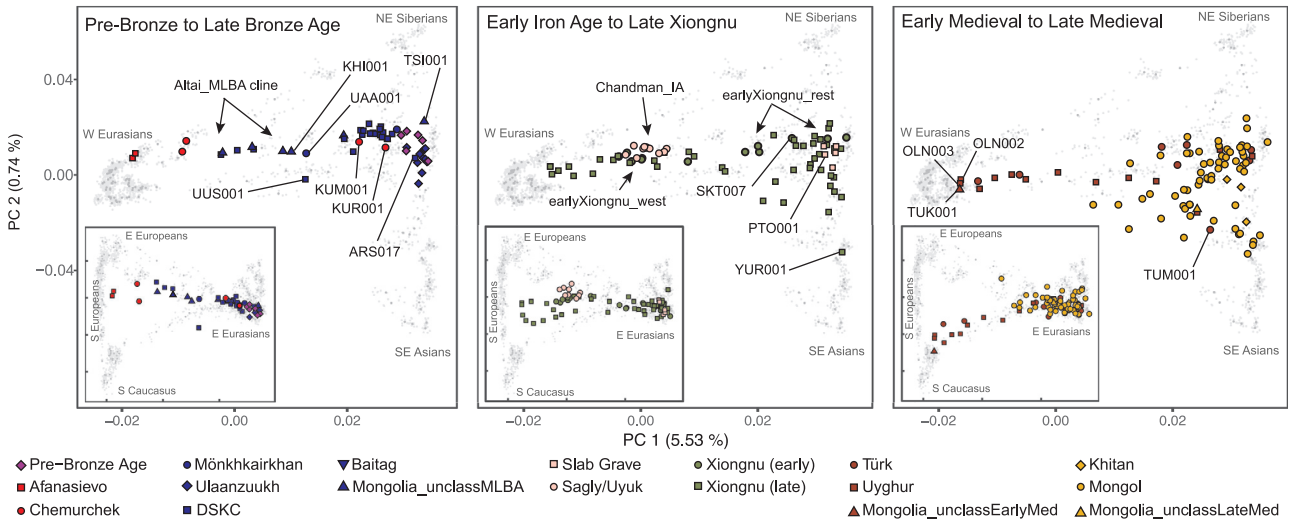


Figure 2. Genetic Structure of Mongolia through Time

PCA of ancient individuals (n = 214) from three major periods projected onto contemporary Eurasians (gray symbols). Main panels display PC1 versus PC2; insets display PC1 versus PC3. Inset tick marks for PC1 correspond to those for the main panels; PC3 accounts for 0.35% of variation. See Figure S3B for population, sample, and axis labels, and Tables S1B, S1C, and S2A for further site and sample details.

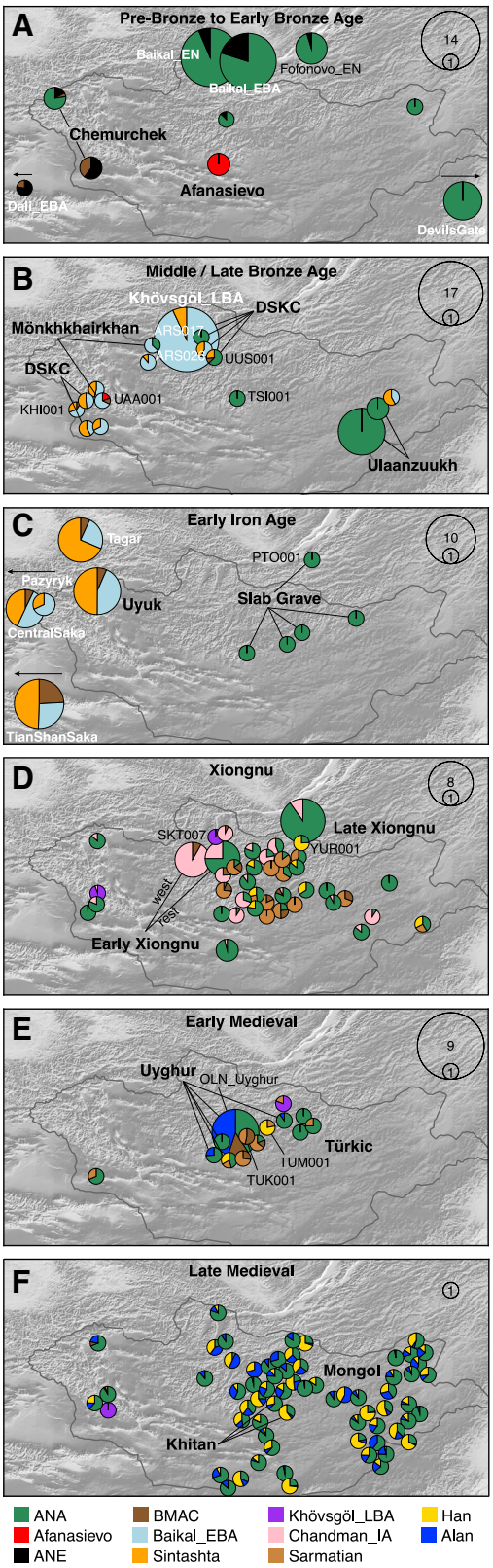
and two individuals from the northern Altai sites of Khundii Gobi (KUM001) and Khuurai Gobi 2 (KUR001). Compared to Afanasievo_Mongolia, the Yagshiin Huduu individuals also show a high degree of Western ancestry but are displaced in PCA (Figure 2) and have a strong genetic affinity with ANE-related ancient individuals such as AfontovaGora3 (AG3), West_Siberia_N, and Botai (Figure 3A; Figures S5A and S5C). We find that the Yagshiin Huduu Chemurchek individuals ("Chemurchek_southAltai") are genetically similar to Dali_EBA (Figure 3A), a contemporaneous individual from eastern Kazakhstan (Narasimhan et al., 2019). The genetic profiles of both the Yagshiin Huduu and Dali_EBA individuals are well fitted by two-way admixture models with Botai (60%–78%) and groups with ancient Iranian-related ancestry, such as Gonur1_BA from Gonur Tepe, a key EBA site of the Bactria-Margiana Archaeological Complex (BMAC) (22%–40%; Figure 3A; Table S5B). Although minor genetic contributions from the Afanasievo-related groups cannot be excluded, Iranian-related ancestry is required for all fitting models, and this admixture is estimated to have occurred 12 ± 6 generations earlier (~336 ± 168 years; Figure S6) when modeled using DATES (Narasimhan et al., 2019). However, because all proxy source populations used in this modeling are quite distant in either time or space from the EBA Altai, the proximate populations contributing to the Chemurchek cannot yet be precisely identified. In the northern Altai, the two Chemurchek individuals ("Chemurchek_northAltai") have mostly ANA-derived ancestry (~80%), with the remainder resembling that of the southern Altai Chemurchek individuals (Figures 3A and 4A; Table S5B). As such, we observe genetic heterogeneity among Chemurchek individuals by geographic location.

Although based on a small number of genomes, we find that neither the Afanasievo nor the Chemurchek left enduring genetic traces into the subsequent MLBA. This is strikingly different than in Europe, where migrating EBA steppe herders had a transformative and lasting genetic impact on local populations (Allentoft et al., 2015; Haak et al., 2015; Mathieson et al., 2018). In the Eastern Steppe, the transient genetic impact of the EBA herders stands in sharp contrast to their strong and enduring cultural and economic impact given that the cultural features that EBA pastoralists first introduced, such as mortuary mound building and dairy pastoralism, continue to the present day.

Bronze Age Emergence of a Tripartite Genetic Structure

Previously, we reported a shared genetic profile among EBA western Baikal hunter-gatherers (Baikal_EBA) and Late Bronze Age (LBA) pastoralists in northern Mongolia (Khövsgöl_LBA) (Jeong et al., 2018). This genetic profile, composed of major and minor ANA and ANE ancestry components, respectively, is also shared with the earlier eastern Baikal (Fofonovo_EN) and Mongolian (centralMongolia_preBA) groups analyzed in this study (Figures 3A, 3B, and 4A), suggesting a regional persistence of this genetic profile for nearly three millennia. Centered in northern Mongolia, this genetic profile is distinct from that of other Bronze Age groups. Overall, we find three distinct and geographically structured gene pools in LBA Mongolia, with the Khövsgöl_LBA population representing one of them (Figures 3B and 4A). The other two, which we refer to as "Altai_MLBA" and "Ulaanzuukh_SlabGrave," are described below.

During the MLBA (1900–900 BCE), as grasslands expanded in response to climate change, new pastoralist cultures expanded out of inner-montane regions and across the Eastern Steppe (Kindstedt and Ser-Od, 2019). This period is also notable for the first regional evidence of horse milking (ca. 1200 BCE; Wilkin et al., 2020a), which is today exclusively associated with alcohol (*airag*) production (Bat-Oyun et al., 2015), and a dramatic intensification of horse use, including the emergence of mounted horseback riding, which would have substantially extended the accessibility of remote regions of the steppe. In the Altai-Sayan



region, dairy pastoralists associated with DSKC and other unclassified MLBA burial types (Altai_MLBA, n = 7) show clear genetic evidence of admixture between a Khövsgöl_LBA-related ancestry and a Sintashta-related WSH ancestry (Figure 3B; Figure S4B). Overall, they form an "Altai_MLBA" cline on PCA between Western Steppe groups and the Baikal_EBA/Khövsgöl_LBA cluster (Figure 2), with their position varying on PC1 according to their level of Western ancestry (Table S5C).

This is the first appearance on the Eastern Steppe of a Sintashta-like ancestry (frequently referred to as "steppe_MLBA" in previous studies), which is distinct from prior Western ancestries present in the Afanasievo and Chemurchek populations and instead shows a close affinity to European Corded-Ware populations and later Andronovo-associated groups, such as the Sintashta (Allentoft et al., 2015). In Khovd province, individuals belonging to DSKC and unclassified MLBA groups (BER002 and SBG001, respectively) have a similar genetic profile that is best modeled as an equal mixture of Khövsgöl_LBA and Sintashta (Figure 3B; Table S5C). This genetic profile matches that previously described for a genetic outlier in northern Mongolia that deviated from the Khövsgöl_LBA cluster in a previous study (ARS026; Jeong et al., 2018). An additional four Altai_MLBA individuals belonging to DSKC (ULI001) and unclassified MLBA groups (BIL001, ULI003, ULZ001) also fit this admixture model with varying admixture proportions (Table S5C). Taken together, the Altai_MLBA cline reveals the ongoing mixture of two source populations: a Sintashta/Andronovo-related WSH population and a local population represented by Khövsgöl_LBA. The admixture is estimated to have occurred only 10 ± 2 generations (~290 years) before the individuals analyzed in this study, a finding consistent with their heterogeneous ancestry proportions (Figure S6). Because the Sintashta culture (ca. 2200–1700 BCE) is associated with novel transportation technologies, such as horse-drawn chariots (Anthony, 2010), the appearance of this ancestry profile on the Eastern Steppe suggests that heightened mobility capabilities played an important role in linking diverse populations across the Eurasian Steppe (Honeychurch, 2015).

Three MLBA individuals in our dataset present genetic profiles that cannot be fully explained by the Altai_MLBA cline. These three, two Altai individuals (UAA001 and KHI001) and UUS001 from Khövsgöl province, are better modeled with a small

Figure 3. Genetic Changes in the Eastern Steppe across Time Characterized by qpAdm

(A–F) Major time periods: (A) Pre-Bronze through Early Bronze Age, (B) Middle/Late Bronze Age, (C) Early Iron Age, (D) Xiongnu period, (E) Early Medieval, and (F) Late Medieval.

Modeled ancestry proportions are indicated by sample size-scaled pie charts, with ancestry source populations shown below (see STAR Methods). The sample size range for each panel is indicated in the upper right. For (B) and (C), Baikal_EBA is modeled as light blue; in (D–F), Khövsgöl_LBA (purple) and the Sagly/Uyuk of Chandman_IA (pink) are modeled as new sources (Figure 4). Cultural groups are indicated by bold text. For (D–F), individuals are Late Xiongnu, Türkic, and Mongol, respectively, unless otherwise noted. Previously published reference populations are noted with white text; all others are from this study. Populations beyond the map borders are indicated by arrows. Burial locations have been jittered to improve visibility of overlapping individuals.

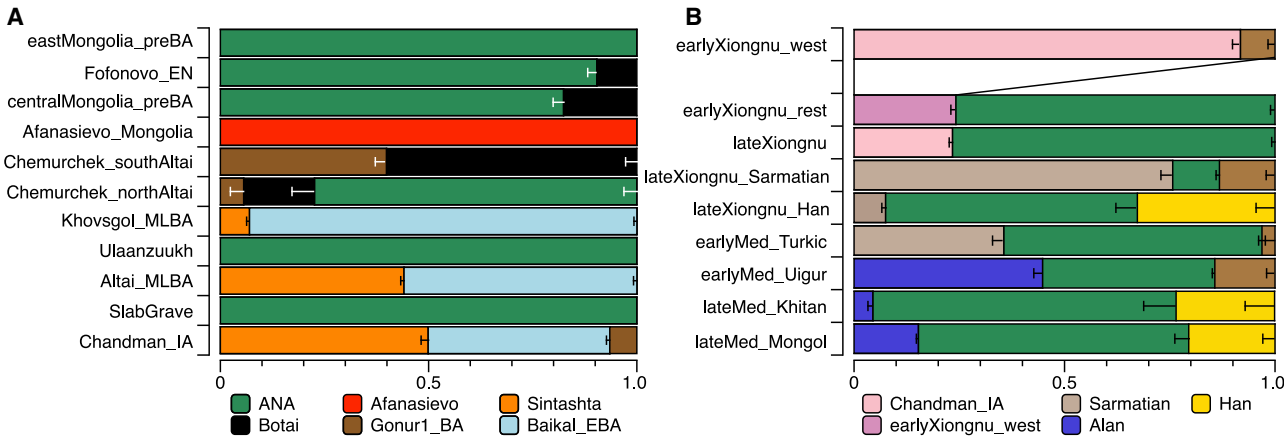


Figure 4. Genetic Ancestry Changes in Chronological Order across All Newly Reported Genetic Groups

Well-fitted modeling results for grouped-based population genetics analyses for (A) prehistoric periods and (B) historic periods. The number of individuals in each genetic group is given in Table S3A. Raw ancestry proportions and standard error estimates are provided in Table S5. Horizontal bars represent ± 1 standard error (SE) estimated by qpAdm.

contribution from Gonur1_BA as a third ancestry source (Table S5C). Taken together, although cultural differences may have existed among the major MLBA mortuary traditions of the Altai and northern Mongolia (Mönkhkhairkhan, DSKC, and unclassified MLBA), they do not form distinct genetic groups.

The populations making up the heterogeneous Altai_MLBA cline left descendants in the Altai-Sayan region, who we later identify at the Sagly/Uyuk site of Chandman Mountain ("Chandman_IA," ca. 400–200 BCE) in northwestern Mongolia during the Early Iron Age (EIA). Nine Chandman_IA individuals form a tight cluster on PCA at the end of the previous Altai_MLBA cline away from Khövsgöl_LBA cluster (Figure 2). During the EIA, the Sagly/Uyuk were pastoralists and millet agropastoralists largely centered in the Upper Yenisei region of present-day Tuva. Together with the Pazyryk of the Altai and the Saka of eastern Kazakhstan, they formed part of a broader Scythian cultural phenomenon that stretched across the Western Steppe, Tarim Basin, and Upper Yenesei (Parzinger, 2006).

We find that EIA Scythian populations systematically deviate from the earlier Altai_MLBA cline, requiring a third ancestral component (Figures 3C and 4A; FigureS4C). The appearance of this ancestry, related to populations of Central Asia (Caucasus/Iranian Plateau/Transoxiana regions) including BMAC (Narasimhan et al., 2019), is clearly detected in the Iron Age groups such as Central Saka, TianShan Saka, Tagar (Damgaard et al., 2018b), and Chandman_IA, while absent in the earlier DSKC and Karasuk groups (Tables S5C–S5E). This third component makes up 6%–24% of the ancestry in these Iron Age groups, and the date of admixture in Chandman_IA is estimated at ~18 ± 4 generations earlier, ca. 750 BCE, which postdates the collapse of the BMAC ca. 1600 BCE and slightly predates the formation of the Persian Achaemenid empire ca. 550 BCE (Figure S6). We suggest that this Iranian-related genetic influx was mediated by increased contact and mixture with agropastoralist populations in the region of Transoxiana (Turan) and Fergana during the LBA to EIA transition. The widespread emergence of horseback riding during the late second and early first millennium BCE (Drews, 2004), and the increasing sophistication of horse transport thereafter, likely contributed to increased population contact and the dissemination of this Iranian-related ancestry onto the steppe. Our results do not exclude additional spheres of contact, such as increased mobility along the Inner Asian Mountain Corridor, which could have also introduced this ancestry into the Altai via Xinjiang starting in the Bronze Age (Frachetti, 2012).

In contrast to the MLBA and EIA cultures of the Altai and northern Mongolia, different burial traditions are found in the eastern and southern regions of Mongolia (Honeychurch, 2015), notably the LBA Ulaanzuukh (1450–1150 BCE) and EIA Slab Grave (1000–300 BCE) cultures. In contrast to other contemporaneous Eastern Steppe populations, we find that individuals associated with these burial types show a clear northeastern Asian (ANA-related) genetic profile lacking both ANE and WSH admixture (Figures 2, 3C, and 4). Both groups were ruminant pastoralists, and the EIA Slab Grave culture also milked horses (Wilkin et al., 2020a). The genetic profiles of Ulaanzuukh and Slab Grave individuals are genetically indistinguishable (Figure 2; Table S5C), consistent with the archaeological hypothesis that the Slab Grave tradition emerged out of the Ulaanzuukh (Honeychurch, 2015; Khatanbaatar, 2019). Both groups are also indistinguishable from the earlier eastMongolia_preBA individual dating to ca. 4600 BCE, suggesting a long-term (>4,000-year) stability of this prehistoric eastern Mongolian gene pool (Table S5C). In subsequent analyses, we merged Ulaanzuukh and Slab Grave into a single genetic group ("Ulaanzuukh_SlabGrave"). The Ulaanzuukh_SlabGrave genetic cluster is the likely source of the previously described DSKC eastern outlier from Khövsgöl province (ARS017) (Jeong et al., 2018), as well as a culturally unclassified individual (TSI001) from central Mongolia who dates to the LBA-EIA transition (Figures 2, 3B, and 3C; Table S5C). In addition, the Mönkhkhairkhan individual KHU001 from northwest Mongolia has a non-negligible amount of Ulaanzuukh_SlabGrave ancestry in addition to his otherwise Baikal_EBA ancestry (Figure S4B; Table S5C). While these three

individuals attest to occasional long-distance contacts between northwestern and eastern Mongolia during the LBA, we find no evidence of Ulaanzuukh_SlabGrave ancestry in the Altai, and the overall frequency of the Ulaanzuukh_SlabGrave genetic profile outside of eastern and southern Mongolia during the MLBA is very low. During the EIA, the Slab Grave culture expanded northward, sometimes disrupting and uprooting former DSKC graves in their path (Fitzhugh, 2009; Honeychurch, 2015; Tsybiktarov, 2003; Volkov, 2002), and it ultimately reached as far north as the eastern Baikal region, which is reflected in the genetic profile of the Slab Grave individual PTO001 in this study (Figure 3C). Overall, our findings reveal a strong east-west genetic division among Bronze Age Eastern Steppe populations through the end of the Early Iron Age. Further sampling from central and southern Mongolia will help refine the spatial distribution of these ancestry profiles, as well as the representativeness of our current findings.

The Xiongnu Empire, the Rise of the First Imperial Steppe Polity

Arising from the prehistoric populations of the Eastern Steppe, large-scale polities began to develop during the late first millennium BCE. The Xiongnu was the first historically documented empire founded by pastoralists, and its establishment is considered a watershed event in the sociopolitical history of the Eastern Steppe (Brosseder and Miller, 2011; Honeychurch, 2015). The Xiongnu held political dominance in East and Central Asia from the third century BCE through the first century CE. The cultural, linguistic, and genetic makeup of the people who constituted the Xiongnu empire has been of great interest, as has their relationship to other contemporaneous and subsequent nomadic groups on the Eastern Steppe. Here, we report genome-wide data for 60 Xiongnu-era individuals from across Mongolia and dating from ca. 200 BCE to 100 CE, thus spanning the entire period of the Xiongnu empire. Although most individuals date to the late Xiongnu period (after 50 BCE), 13 individuals predate 100 BCE and include 12 individuals from the northern early Xiongnu frontier sites of Salkhityn Am (SKT) and Atsyn Gol (AST) and one individual from the early Xiongnu site of Jargalantyn Am (JAG) in eastern Mongolia.

We observe two distinct demographic processes that contributed to the formation of the early Xiongnu. First, half of the early individuals (n = 6) form a genetic cluster (earlyXiongnu_west) resembling that of Chandman_IA of the preceding Sagly/Uyuk culture from the Altai-Sayan region (Figure 2). They derive 92% of their ancestry from Chandman_IA with the remainder attributed to additional Iranian-related ancestry, which we model using BMAC as a proxy (Figures 3D and 4D; Table S5F). This suggests that the low-level Iranian-related gene flow identified among the Chandman_IA Sagly/Uyuk during the EIA likely continued during the second half of the first millennium BCE, spreading across western and northern Mongolia. Second, six individuals (“earlyXiongnu_rest”) fall intermediate between the earlyXiongnu_west and Ulaanzuukh_SlabGrave clusters; four carry varying degrees of earlyXiongnu_west (39%–75%) and Ulaanzuukh_SlabGrave (25%–61%) related ancestry, and two (SKT004, JAG001) are indistinguishable from the Ulaanzuukh_SlabGrave cluster (Figure 3D; Tables S5F and S5G). This genetic cline linking the earlyXiongnu_west and Ulaanzuukh_SlabGrave gene pools signifies the unification of two deeply diverged and distinct lineages on the Eastern Steppe—between the descendants of the DSKC, Mönkhkhairkhan, and Sagly/Uyuk cultures in the west and the descendants of the Ulaanzuukh and Slab Grave cultures in the east. Overall, the low-level influx of Iranian-related gene flow continuing from the previous Sagly/Uyuk culture and the sudden appearance of a novel east-west mixture uniting the gene pools of the Eastern Steppe are the two defining demographic processes associated with the rise of the Xiongnu.

Among late Xiongnu individuals, we find even higher genetic heterogeneity (Figure 2), and their distribution on PC indicates that the two demographic processes evident among the early Xiongnu continued into the late Xiongnu period, but with the addition of new waves and complex directions of gene flow. Of the 47 late Xiongnu individuals, half (n = 26) can be adequately modeled by the same admixture processes seen among the early Xiongnu: 22 as a mixture of Chandman_IA+Ulaanzuukh_SlabGrave, 2 (NAI002, TUK002) as a mixture of either Chandman_IA+BMAC or Chandman_IA+Ulaanzuukh_SlabGrave+BMAC, and 2 (TUK003, TAK001) as a mixture of either earlyXiongnu_west+Ulaanzuukh_SlabGrave or earlyXiongnu_west+Khovsgol_LBA (Figures 3D and 4D; Table S5G). A further two individuals (TEV002, BUR001) also likely derive their ancestry from the early Xiongnu gene pool, although the p value of their models is slightly lower than the 0.05 threshold (Table S5G). However, a further 11 late Xiongnu with the highest proportions of western Eurasian affinity along PC1 cannot be modeled using BMAC or any other ancient Iranian-related population. Instead, they fall on a cluster of ancient Sarmatians from various locations in the Western and Central Steppe (Figure 2).

Admixture modeling confirms the presence of a Sarmatian-related gene pool among the late Xiongnu: three individuals (UGU010, TMI001, BUR003) are indistinguishable from Sarmatian, two individuals (DUU001, BUR002) are admixed between Sarmatian and BMAC, three individuals (UGU005, UGU006, BRL002) are admixed between Sarmatian and Ulaanzuukh_SlabGrave, and three individuals (NAI001, BUR004, HUD001) require Sarmatian, BMAC, and Ulaanzuukh_SlabGrave (Figure 3D; Figure S4D; Table S5G). In addition, eight individuals with the highest eastern Eurasian affinity along PC1 are distinct from both the Ulaanzuukh_SlabGrave and Khövsgöl_LBA genetic profiles, showing affinity along PC2 toward present-day people from East Asia further to the south (Figure 2). Six of these individuals (EME002, ATS001, BAM001, SON001, TUH001, YUR001) are adequately modeled as a mixture of Ulaanzuukh_SlabGrave and Han (Tables S5F and S5G), and YUR001 in particular exhibits a close genetic similarity to two previously published Han empire soldiers (Damgaard et al., 2018b), whose genetic profile we refer to as “Han_2000BP” (Table S5G). The remaining two individuals (BRU001, TUH002) are similar but also require the addition of Sarmatian ancestry (Table S5G). The late Xiongnu are thus characterized by two additional demographic processes that distinguish them from the early Xiongnu: gene flow from a new Sarmatian-related Western ancestry source and intensified interaction and mixture with people of the contemporaneous Han empire of China. A previous study

of the Egyin Gol Xiongnu necropolis reported mitochondrial haplogroups of both western and eastern Eurasian origins (Keyser-Tracqui et al., 2003), and this accords with our findings of the west-east admixture from genome-wide data. Together, these results match well with historical records documenting the political influence that the Xiongnu exercised over their neighbors, including the Silk Road kingdoms of Central Asia and Han Dynasty China, as well as purported migrations both in and out of Mongolia (Miller, 2014). Overall, the Xiongnu period can be characterized as one of expansive and extensive gene flow that began by uniting the gene pools of western and eastern Mongolia and ended by uniting the gene pools of western and eastern Asia.

Fluctuating Genetic Heterogeneity in the Post-Xiongnu Polities

After the collapse of the Xiongnu empire ca. 100 CE, a succession of nomadic pastoralist regimes rose and fell over the next several centuries across the politically fragmented Eastern Steppe: Xianbei (ca. 100–250 CE), Rouran (ca. 300–550 CE), Türkic (552–742 CE), and Uyghur (744–840 CE). Although our sample representation for the Early Medieval period is uneven, consisting of 1 unclassified individual dating to the Xianbei or Rouran period (TUK001), 8 individuals from Türkic mortuary contexts, and 13 individuals from Uyghur cemeteries, it is clear that these individuals have genetic profiles that differ from the preceding Xiongnu period, suggesting new sources of gene flow into Mongolia at this time that displace them along PC3 (Figure 2). Individual TUK001 (250–383 cal. CE), whose burial was an intrusion into an earlier Xiongnu cemetery, has the highest western Eurasian affinity. This ancestry is distinct from that of the Sarmatians and closer to ancient populations with BMAC/Iranian-related ancestry (Figure 2). Among the individuals with the highest eastern Eurasian affinity, two Türkic-period individuals and one Uyghur-period individual (ZAA004, ZAA002, OLN001.B) are indistinguishable from the Ulaanzuukh_SlabGrave cluster. Another individual (TUM001), who was recovered from the tomb ramp of an elite Türkic-era emissary of the Tang Dynasty, has a high proportion of Han-related ancestry (78%; Figures 3E and 4B; Figure S4E; Table S5H). This male, buried with two dogs, was likely a Chinese attendant sacrificed to guard the tomb entrance (Ochir et al., 2013). The remaining 17 Türkic and Uyghur individuals show intermediate genetic profiles (Figure 3E).

The high genetic heterogeneity of the Early Medieval period is vividly exemplified by 12 individuals from the Uyghur period cemetery of Olon Dov (OLN; Figure 2) in the vicinity of the Uyghur capital of Ordu-Baliq. Six of these individuals came from a single tomb (grave 19), of whom only two are related (OLN002 and OLN003, second-degree; Table S2D); the absence of closer kinship ties raises questions about the function of such tombs and the social relationships of those buried within them. Most Uyghur-period individuals exhibit a high but variable degree of west Eurasian ancestry—best modeled as a mixture of Alans, a historic nomadic pastoral group likely descended from the Sarmatians and contemporaries of the Huns (Bachrach, 1973), and an Iranian-related (BMAC-related) ancestry—together with Ulaanzuukh_SlabGrave (ANA-related) ancestry (Figure 3E). The admixture dates estimated for the ancient Türkic and Uyghur individuals in this study correspond to ca. 500 CE: 8 ± 2 generations before the Türkic individuals and 12 ± 2 generations before the Uyghur individuals (represented by ZAA001 and Olon Dov individuals).

Rise of the Mongol Empire

After the fall of the Uyghur empire in the mid-ninth century, the Khitans of northeast China established the powerful Liao Dynasty in 916 CE. The Khitans controlled large areas of the Eastern Steppe and are recorded to have relocated people within their conquered territories (Kradin and Ivliev, 2008), but few Khitan period cemeteries are known within Mongolia. Our study includes three Khitan individuals (ZAA003, ZAA005, ULA001) from Bulgan province, all of whom have a strongly eastern Eurasian genetic profile (Figure 2), with <10% west Eurasian ancestry (Figures 3F and 4B; Table S5I). This may reflect the northeastern Asian origin of the Mongolic-speaking Khitan, but a larger sample size is required to adequately characterize the genetic profile of Khitan populations within Mongolia. In 1125 CE, the Khitan empire fell to the Jurchen's Jin Dynasty, which was then conquered in turn by the Mongols in 1234 CE.

At its greatest extent, the Mongol empire (1206–1368 CE) spanned nearly two-thirds of the Eurasian continent. It was the world's largest contiguous land empire, and the cosmopolitan entity comprised diverse populations that flowed into the steppe heartland. We analyzed 62 Mongol-era individuals whose burials are consistent with those of low-level, local elites. No royal or regional elite burials were included, and neither were individuals from the cosmopolitan capital of Karakorum. Although we find that Mongol-era individuals were diverse, they exhibit a much lower genetic heterogeneity than the Xiongnu-era individuals (Figure 2), and they almost entirely lack the residual ANE-related ancestry (in the form of Chandman_IA and Khövsgöl_LBA) that had been present among the Xiongnu and earlier northern/western MLBA cultures. On average, Mongol-period individuals have a much higher eastern Eurasian affinity than previous empires, and this period marks the beginning of the formation of the modern Mongolian gene pool. We find that most historic Mongols are well-fitted by a three-way admixture model with the following ancestry proxies: Ulaanzuukh_SlabGrave, Han, and Alans. Consistent with their PCA location (Figure 2), Mongol-era individuals as a group can be modeled with only 15%–18% Western Steppe ancestry (Alan or Sarmatian) but require 55%–64% Ulaanzuukh_SlabGrave and 21%–27% of Han-related ancestry (Table S5I). Applying the same model to each individual separately, this three-source model adequately explains 56 out of 61 ancient Mongols (based on p value at threshold of 0.05), as well as one unclassified Late Medieval individual dating to around the beginning of the Mongol empire (SHU002) (Table S5J).

Since the fall of the Mongol empire in 1368 CE, the genetic profile of the Mongolian populations has not substantially changed. The genetic structure established during the Mongol empire continues to characterize present-day Mongolic-speaking populations living in both Mongolia and Russia. We examined the genetic cladality between the historic Mongols and seven present-day Mongolic-speaking groups (Mongols, Kalmyk, Buryat,

Khamnegan, Daur, Tu, and Mongola) using an individual-based qpWave analysis. Within the resolution of current data, 34 of 61 historic Mongols are genetically cladal with at least one modern Mongolic-speaking population (Figure S7B). The Mongol empire had a profound impact on restructuring the political and genetic landscape of the Eastern Steppe, and these effects endured long after the decline of the empire and are still evident in Mongolia today.

Functional and Gendered Aspects of Recurrent Admixture in the Eastern Steppe

To investigate the functional aspects of recurrent admixture on the Eastern Steppe, we estimated the population allele frequency of five SNPs associated with functional or evolutionary aspects of lactose digestion (*LCT/MCM6*), dental morphology (*EDAR*), pigmentation (*OCA2*, *SLC24A5*), and alcohol metabolism (*ADH1B*) (Figure 5A). First, we find that despite a pastoralist lifestyle with widespread direct evidence for milk consumption (Jeong et al., 2018; Wilkin et al., 2020a), the MLBA and EIA individuals of the Eastern Steppe did not have any derived mutations conferring lactase persistence. Individuals from subsequent periods did have the derived mutation that is today widespread in Europe (rs4988235) but at negligibly low frequency (~5%) and with no increase in frequency over time (Figure 5A). This is somewhat remarkable given that, in addition to other dairy products, some contemporary Mongolian herders consume up to 4–10 L of *airag* (fermented mare's milk, ~2.5% lactose) per day during the summer months (Bat-Oyun et al., 2015), resulting in a daily intake of 100–250 g of lactose sugar. Petroglyph depictions of *airag* production date back to the EIA in the Yenisei Basin (Dèvlet, 1976), and accounts of the historic Mongols record abundant and frequent consumption of *airag*, as well as a wide range of additional liquid and solid ruminant dairy products (Bayarsaikhan, 2016; Onon, 2005), which has been additionally confirmed by ancient proteomic evidence (Jeong et al., 2018; Wilkin et al., 2020a). How Mongolians have been able to digest such large quantities of lactose for millennia in the absence of lactase persistence is unknown, but it may be related to their reportedly unusual gut microbiome structure, which today is highly enriched in lactose-digesting *Bifidobacterium* spp. (Liu et al., 2016).

Genetic markers that underwent regional selective sweeps show allele frequency changes that correlate with changes in the genome-wide ancestry profile (Figure 5A). For example, rs3827760 in *EDAR* (ectodysplasin A receptor) and rs1426654 in *SLC24A5* (solute carrier family 24 member 5) are well-known targets of positive selection in East Asians and western Eurasians, respectively (Sabeti et al., 2007). Our MLBA and EIA populations show a strong population differentiation in the allele frequencies of these two SNPs: rs3827760 frequency is much higher in groups with higher eastern Eurasian affinity (Khovsgol_LBA, Ulaanzuukh_SlabGrave), whereas rs1426654 is higher in Altai_MLBA and Chandman_IA (Table S2E). We find that two SNPs that have undergone more recent positive selection (Donnelly et al., 2012; Li et al., 2011) in East Asians, rs1229984 in *ADH1B* (aldehyde dehydrogenase 1B) and rs1800414 in *OCA2* (oculocutaneous albinism II), were absent or in extremely low frequency during the MLBA and EIA, when the eastern Eurasian ancestry was primarily ANA-related, but increased in frequency over time as the proportion of East Asian ancestry increased through interactions with imperial China and other groups (Table S2E).

Finally, we investigated gendered dimensions of the population history of the Eastern Steppe. Sex-biased patterns of genetic admixture can be informative about gendered aspects of migration, social kinship, and family structure. We observe a clear signal of male-biased WSH admixture among the EIA Sagly/Uyuk and during the Türkic period (i.e., more positive *Z* scores; Figure 5B), which also corresponds to the decline in the Y chromosome lineage Q1a and the concomitant rise of the western Eurasian lineages such as R and J (Figure S2A). During the later Khitan and Mongol empires, we observe a prominent male bias for East Asian-related ancestry (Figure S2C), which can also be seen from the rise in frequency of Y chromosome lineage O2a (Figure S2A). The Xiongnu period exhibits the most complex pattern of male-biased admixture, whereby different genetic subsets of the population exhibit evidence of different sources of male-biased admixture (Figure S2C).

Among the Xiongnu, we also detect 10 genetic relative pairs, including a father-daughter pair buried in the same grave (JAG001 and JAA001) at Jargalantyn Am, as well as a mother-son pair (IMA002 and IMA005) at Il'movaya Pad, a brother-sister pair (TMI001 and BUR003) at Tamiryn Ulaan Khoshuu, and a brother-brother pair (SKT002 and SKT006) at Salkhityn Am (Table S2D). Of the remaining six pairs, three are female-female relative pairs buried within the same site, suggesting the presence of extended female kinship within Xiongnu groups. First-degree relatives within a single site have also been reported in a previous study on the Egyin Gol Xiongnu necropolis based on the autosomal short tandem repeat (STR) data (Keyser-Tracqui et al., 2003). These relationships, when combined with mortuary features, offer the first clues to local lineage and kinship structures within the Xiongnu empire, which are otherwise poorly understood.

DISCUSSION

The population history of the Eastern Steppe is one marked by the repeated mixing of diverse eastern and western Eurasian gene pools. However, rather than simple waves of migration, demographic events on the Eastern Steppe have been complex and variable. Generating more than 200 genome-wide ancient datasets, we have presented the first genetic evidence of this dynamic population history, from ca. 4600 BCE through the end of the Mongol empire. We found that the Eastern Steppe was populated by hunter-gatherers of ANA and ANE ancestry during the mid-Holocene and then shifted to a dairy pastoralist economy during the Bronze Age. Migrating Yamnaya/Afanasievo steppe herders, equipped with carts and domestic livestock (Kovalev and Erdenebaatar, 2009), appear to have first introduced ruminant dairy pastoralism ca. 3000 BCE (Wilkin et al., 2020a) but surprisingly had little lasting genetic impact, unlike in Europe (Allentoft et al., 2015; Haak et al., 2015; Mathieson et al., 2015). By the MLBA, ruminant dairy pastoralism had been adopted by populations throughout the Eastern Steppe (Wilkin et al., 2020a), regardless of ancestry, and this subsistence has

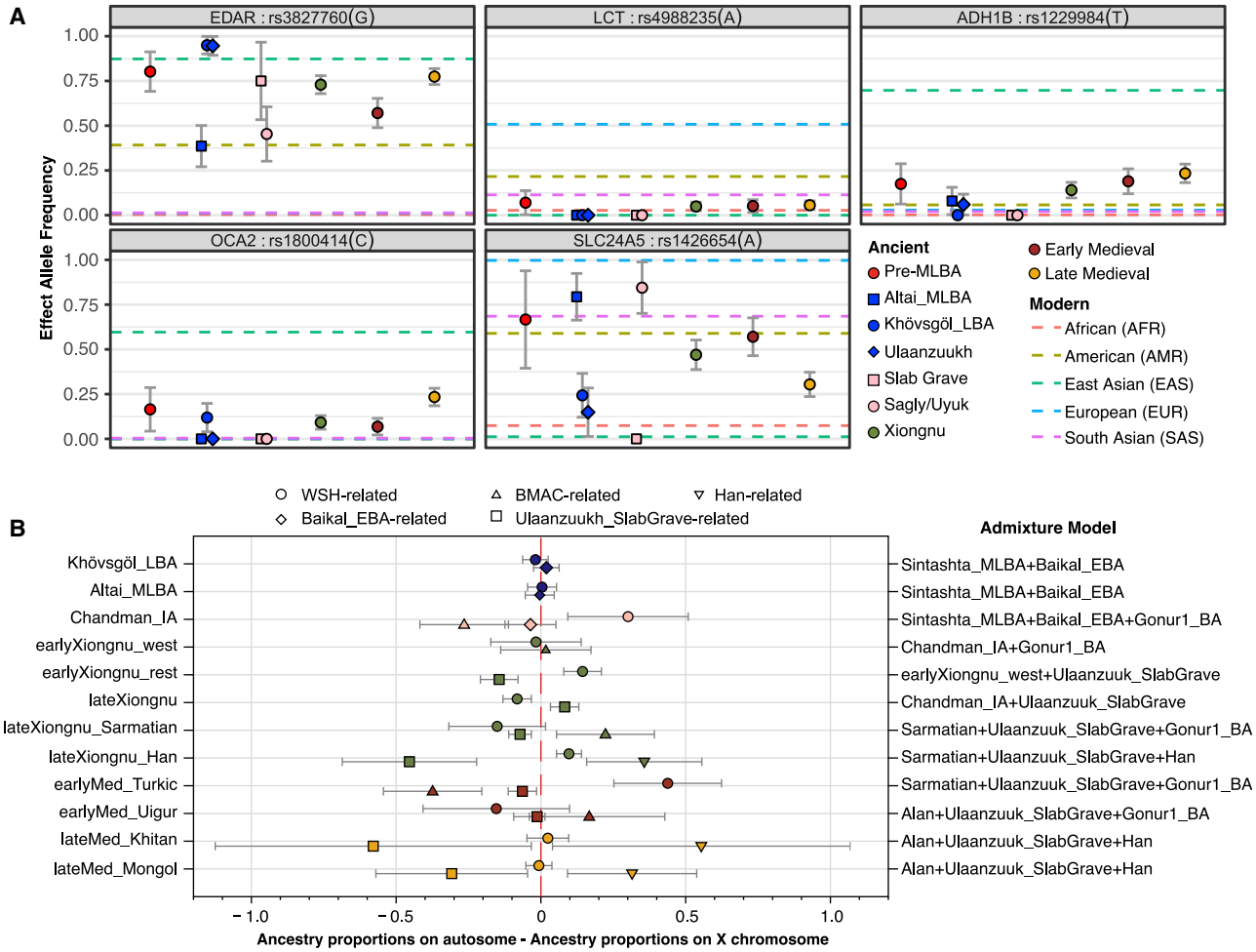


Figure 5. Functional Allele Frequencies and Sex-Biased Patterns of Genetic Admixture

(A) Allele frequencies of five phenotypic SNP changes through time. For the effective allele, we show maximum likelihood frequency estimates and one standard error bar for each ancient group. The pre-MLBA category corresponds to the sum of all ancient groups before Mönkhkhairkhan. Xiongnu, Early Medieval, and Late Medieval correspond to the sum of all ancient groups in each period correspondingly. Horizontal dashed lines show allele frequency information from the 1000 Genomes Project's five super populations.

(B) Sex-biased patterns of genetic admixture by period and population. We calculated *Z* scores for every ancient individual who has genetic admixture with WSH-/Iranian-/Han-related ancestry. Positive scores suggest more WSH-/Iranian-/Han-related ancestry on the autosomes, i.e., male-driven admixture. See Figure S2C for individual *Z* scores.

continued, with the additions of horse milking in the LBA and camel milking in the Mongol period (Wilkin et al., 2020a), to the present day (Bat-Oyun et al., 2015; Kindstedt and Ser-Od, 2019). Puzzlingly, however, there is no evidence of selection for lactase persistence over this 5,000-year history, despite the repeated introduction of this genetic trait by subsequent migrations of groups from the west. This suggests a different trajectory of lactose adaptation in Asia that to date remains unexplained.

During the MLBA, we observed the formation of a tripartite genetic structure on the Eastern Steppe, characterized by the continuation of pre-Bronze Age ANA ancestry in the east and a cline of genetic variation between pre-Bronze Age ANA-ANE ancestry in the north and increasing proportions of a new Sintashta-related WSH ancestry in the west. The Sintashta, a western forest steppe culture with genetic links to the European Corded Ware cultures (Mathieson et al., 2015), were masters of bronze metallurgy and chariotry (Anthony, 2010), and the appearance of this ancestry on the Eastern Steppe may be linked to the introduction of new (especially horse-related) technologies. DSKC sites in particular show widespread evidence for horse use in transport and perhaps even riding (Taylor et al., 2015), and genetic analysis has demonstrated a close link between these animals and the Sintashta chariot horses (Fages et al., 2019). The strong east-west genetic division among Bronze Age Eastern Steppe populations at this time was maintained for more than a millennium and through the end of the EIA, when the first clear evidence for widespread horseback riding appears (Drews, 2004) and the heightened mobility of some groups, notably the eastern Slab Grave culture (Honeychurch, 2015), began to disrupt this structure. Eventually, the

three major ancestries met and mixed, and this was contemporaneous with the emergence of the Xiongnu empire. The Xiongnu are characterized by extreme levels of genetic heterogeneity and increased diversity as new and additional ancestries from China, Central Asia, and the Western Steppe (Sarmatian-related) rapidly entered the gene pool.

Genetic data for the subsequent Early Medieval period are relatively sparse and uneven, and few Xianbei or Rouran sites have yet been identified during the 400-year gap between the Xiongnu and Türkic periods. We observed high genetic heterogeneity and diversity during the Türkic and Uyghur periods, and following the collapse of the Uyghur empire, we documented a final major genetic shift during the late medieval period toward greater eastern Eurasian ancestry, which is consistent with historically documented expansions of Tungusic- (Jurchen) and Mongolic- (Khitan and Mongol) speaking groups from the northeast into the Eastern Steppe (Biran, 2012). We also observed that this East Asian-related ancestry was brought into the Late Medieval populations more by male than female ancestors. By the end of the Mongol period, the genetic makeup of the Eastern Steppe had dramatically changed, retaining little of the ANE ancestry that had been a prominent feature during its prehistory. Today, ANE ancestry survives in appreciable amounts only in isolated Siberian groups and among the indigenous peoples of the Americas (Jeong et al., 2019). The genetic profile of the historic Mongols is still reflected among contemporary Mongolians, suggesting a relative stability of this gene pool over the last ~700 years.

Having documented key periods of genetic shifts in the Eastern steppe, future work may be able to explore whether these shifts are also linked to cultural and technological innovations and how these innovations may have influenced the political landscape. Integrating these findings with research on changes in horse technology and herding practices, as well as shifts in livestock traits and breeds, may prove particularly illuminating. This study represents the first large-scale paleogenomic investigation of the Eastern Eurasian Steppe, and it sheds light on the remarkably complex and dynamic genetic diversity of the region. Despite this progress, there is still a great need for further genetic research in central and eastern Eurasia, and particularly in northeastern China, the Tarim Basin, and the eastern Kazakh steppe, in order to fully reveal the population history of the Eurasian Steppe and its pivotal role in world prehistory.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2020.10.015.

ACKNOWLEDGMENTS

We thank D. Navaan, Z. Batsaikhan, V. Volkow, D. Tseveendorj, L. Erdenebold, M. Horton, Sh. Uranchimeg, B. Naran, N. Ser-Odjav, P. Konovalov, Kh. Lhagvasuren, N. Mamonova, E. Mijiddorj, L. Namsrainaidan, A.P. Okladnikov, Kh. Perlee, S. Danilov, and D.I. Burayev for their archaeological contributions and Solodovnikov Konstantin for his physical anthropological analysis. We thank the Department of Archaeology and Physical Anthropology at the National University of Mongolia and the Museum of the Buryat Scientific Center, SB RAS for their cooperation on this project. We thank R. Flad for comments on early manuscript drafts, M. Bleasdale and S. Gankhuyg for assistance with sampling, S. Nagel and M. Meyer for assistance with ssDNA library preparation, S. Nakagome for sharing a script to estimate population allele frequency from low-coverage sequence data, and P. Moorjani for granting access to the DATES program prior to publication. Permissions: The human remains analyzed in this study were reviewed and approved by the Mongolian Ministry of Culture and the Mongolian Ministry of Education, Culture, Science, and Sport under reference numbers A0122772 MN DE 0 8124, A0109258 MN DE 7 643, and A0117901 MN DE 9 4314 and declaration number 12-2091008-20E00225. Funding: This research was supported by the Max Planck Society, the Ministry of Education, Culture, Science and Sport of Mongolia (grant #2018/25 to E.M. and D.T.), the Russian Foundation for Basic Research (grant #18-59-94020 to E.M. and D.T.), the Ministry of Education and Science of the Russian Federation (grant #14.W03.31.0016 to N.K., B.A.B., D.A.M., and P.B.K.), the US National Science Foundation (BCS-1523264 to C.W.), the Deutsche Forschungsgemeinschaft (SFB 1167 no. 257731206 to J.B.), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2020R1C1C1003879 to C.J.), and the European Research Council under the European Union's Horizon 2020 research and innovation program (under grant agreement numbers 771234-PALEORIDER to W.H. and 804884-DAIRYCULTURES to C.W.).

AUTHOR CONTRIBUTIONS

C.W., C.J., E.M., and N.B. designed the research; C.W. and C.J. supervised the research; E.M., S.W., J.H., J.H.B., S.U., W.H., N.K., B.A.B., D.A.M, P.B.K., E.Z., A.V.M., N.B., and C.W. provided materials and resources; B.K.M., W.T.T.T., J.H.B., and E.M. performed archaeological data analysis; R.S. and C.C. performed genetic laboratory work; K.W., C.J., F.K., and S.S. performed genetic data analysis; B.K.M., K.W., W.T.T.T, C.J., and C.W.

integrated the archaeological and genetic data; K.W., C.J., and C.W. wrote the paper, with contributions from B.K.M., W.T.T.T., J.H.B., and the other coauthors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 8, 2020
Revised: March 12, 2020
Accepted: October 7, 2020
Published: November 5, 2020

REFERENCES

Allentoft, M.E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L., et al. (2015). Population genomics of Bronze Age Eurasia. Nature *522*, 167–172.

Allsen, T.T. (2015). Population Movements in Mongol Eurasia. In Nomads as Agents of Cultural Change: The Mongols and Their Eurasian Predecessors, R. Amitai and M. Biran, eds. (Honolulu: University of Hawai'i Press), pp. 119–151.

Anthony, D.W. (2010). The Horse, the Wheel, and Language: How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World (Princeton University Press).

Bachrach, B.S. (1973). A History of the Alans in the West (U of Minnesota Press).

Bat-Oyun, T., Erdenetsetseg, B., Shinoda, M., Ozaki, T., and Morinaga, Y. (2015). Who is Making Airag (Fermented Mare's Milk)? A Nationwide Survey of Traditional Food in Mongolia. Nomad. People. *19*, 7–29.

Bayarsaikhan, D. (2016). Drinking Traits and Culture of the Imperial Mongols in the Eyes of Observers and in a Multicultural Context. Crossroads *14*, 161–172.

Beckwith, C.I. (2009). Empires of the Silk Road: A History of Central Eurasia from the Bronze Age to the Present (Princeton University Press).

Bemmann, J., and Nomguunsüren, G. (2012). Bestattungen in Felsspalten und Hohlräumen mongolischer Hochgebirge. In Steppenkrieger: Reiternomaden des 7.-14. Jahrhunderts aus der Mongolei, J. Bemmann, ed. (Darmstadt: WBG), pp. 198–217.

Biran, M. (2012). Kitan migrations in Eurasia (10th–14th centuries). Journal of Central Eurasian Studies *3*, 85–108.

Brosseder, U., and Miller, B.K. (2011). Xiongnu Archaeology: Multidisciplinary Perspectives of the First Steppe Empire in Inner Asia (Vor- und Frühgeschichtliche Archäologie, Rheinische Friedrich-Wilhelms-Universität Bonn).

Clark, J. (2015). Modeling Late Prehistoric and Early Historic Pastoral Adaptations in Northern Mongolia's Darkhad Depression. PhD thesis (University of Pittsburgh).

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., and Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc. Natl. Acad. Sci. USA *110*, 15758–15763.

Damgaard, P.B., Martiniano, R., Kamm, J., Moreno-Mayar, J.V., Kroonen, G., Peyrot, M., Barjamovic, G., Rasmussen, S., Zacho, C., Baimukhanov, N., et al. (2018a). The first horse herders and the impact of early Bronze Age steppe expansions into Asia. Science *360*, eaar7711.

Damgaard, P.B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., Moreno-Mayar, J.V., Pedersen, M.W., Goldberg, A., Usmanova, E., et al. (2018b). 137 ancient human genomes from across the Eurasian steppes. Nature *557*, 369–374.

Dashtseveg, T., Dorjpurev, K., and Myagmar, E. (2014). Bronze Age graves in the Delgerkhaan mountain area of eastern Mongolia and the Ulaanzuukh culture. Asian Archaeol *2*, 40–49.

Devièse, T., Massilani, D., Yi, S., Comeskey, D., Nagel, S., Nickel, B., Ribechini, E., Lee, J., Tseveendorj, D., Gunchinsuren, B., et al. (2019). Compound-specific radiocarbon dating and mitochondrial DNA analysis of the Pleistocene hominin from Salkhit Mongolia. Nat. Commun. *10*, 274.

Dèvlet, M.A. (1976). Bol'shaia Boiarskaia pisanitsa [Rock Engravings in the Middle Yenisei Basin]. (Nauka).

Di Cosmo, N. (2002). Ancient China and Its Enemies: the rise of nomadic power in East Asian history (Cambridge, UK: Cambridge University Press).

Donnelly, M.P., Paschou, P., Grigorenko, E., Gurwitz, D., Barta, C., Lu, R.-B., Zhukova, O.V., Kim, J.-J., Siniscalco, M., New, M., et al. (2012). A global view of the *OCA2-HERC2* region and pigmentation. Hum. Genet. *131*, 683–696.

Dorj, D. (1969). Dornod Mongoliin Neolitiin Ueiin Bulsh, Suutz. Erdem Shinjlgeenii Buteeliin Emkhegtel *1*, 59–70.

Dorjgotiv, D. (2004). Geographic Atlas of Mongolia (Ulaanbaatar: Administration of Land Affairs, Geodesy and Cartography).

Drews, R. (2004). Early Riders: The Beginnings of Mounted Warfare in Asia and Europe (Routledge).

Erdenebaatar, D. (2016). Graves of the Mönkhkhairkhan culture. In Ancient Funeral Monuments of Mongolia, D. Eregzen, ed. (Institute of History and Archaeology, Mongolian Academy of Sciences), pp. 46–57.

Erdenebat, U. (2009). Altmongolisches Grabbrauchtum: archäologisch-historische Untersuchungen zu den mongolischen Grabfunden des 11. bis 17. Jahrhunderts in der Mongolei. PhD thesis (Rheinische Friedrich-Wilhelms-Universität).

Erdenebat, U. (2016). Small Graves of the Uyghur Period. In Ancient Funeral Monuments of Mongolia, G. Eregzen, ed. (Mongolian Academy of Sciences), pp. 230–233.

Erdenebat, U., Batsaikhan, Z., and Dashdorj, B. (2012). Arkhangai aimagiin Khotont sum nutag Olon Dovd 2011 ond yavuulsan arkheologiin sudalgaa. Arkheologiin Sudlal *32*, 229–258.

Eregzen, G. (2011). Treasures of the Xiongnu (Ulaanbaatar: Institute of Archaeology, Mongolian Academy of Sciences).

Eregzen, G. (2016). Ancient funeral monuments of Mongolia (Ulaanbaatar: Institute Of History And Archaeology, Mongolian Academy Of Sciences).

Fages, A., Hanghøj, K., Khan, N., Gaunitz, C., Seguin-Orlando, A., Leonardi, M., McCrory Constantz, C., Gamba, C., Al-Rasheid, K.A.S., Albizuri, S., et al. (2019). Tracking Five Millennia of Horse Management with Extensive Ancient Genome Time Series. Cell *177*, 1419–1435.e31.

Fitzhugh, W.W. (2009). Stone Shamans and Flying Deer of Northern Mongolia: Deer Goddess of Siberia or Chimera of the Steppe? Arctic Anthropol. *46*, 72–88.

Frachetti, M.D. (2012). Multiregional Emergence of Mobile Pastoralism and Nonuniform Institutional Complexity across Eurasia. Curr. Anthropol. *53*, 2–38.

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., Bondarev, A.A., Johnson, P.L.F., Aximu-Petri, A., Prüfer, K., de Filippo, C., et al. (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. Nature *514*, 445–449.

Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., Furtwängler, A., Haak, W., Meyer, M., Mittnik, A., et al. (2016). The genetic history of Ice Age Europe. Nature *534*, 200–205.

Golden, P.B. (1992). An introduction to the history of the Turkic peoples: Ethnogenesis and State-Formation in Medieval and Early Modern Eurasia and the Middle East. (O. Harrassowitz).

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. Nature *522*, 207–211.

Haber, M., Doumet-Serhal, C., Scheib, C., Xue, Y., Danecek, P., Mezzavilla, M., Youhanna, S., Martiniano, R., Prado-Martinez, J., Szpak, M., et al. (2017). Continuity and Admixture in the Last Five Millennia of Levantine History from Ancient Canaanite and Present-Day Lebanese Genome Sequences. Am. J. Hum. Genet. *101*, 274–282.

Hanks, B. (2010). Archaeology of the Eurasian Steppes and Mongolia. Annual Review of Anthropology *39*, 469–486.

Harney, É., May, H., Shalem, D., Rohland, N., Mallick, S., Lazaridis, I., Sarig, R., Stewardson, K., Nordenfelt, S., Patterson, N., et al. (2018). Ancient DNA from Chalcolithic Israel reveals the role of population mixture in cultural transformation. Nat. Commun. *9*, 3336.

Honeychurch, W. (2015). Inner Asia and the Spatial Politics of Empire: Archaeology, Mobility, and Culture Contact (New York, NY: Springer New York).

Honeychurch, W. (2017). The Development of Cultural and Social Complexity in Mongolia. In Handbook of East and Southeast Asian Archaeology, J. Habu, P.V. Lape, and J.W. Olsen, eds. (New York, NY: Springer New York), pp. 513–532.

Hulsewé, A.F.P. (1979). China in Central Asia, the Early Stage: 125 B.C.-A.D. 23 (Leiden: E.J. Brill).

Janz, L., Odsuren, D., and Bukhchuluun, D. (2017). Transitions in Palaeoecology and Technology: Hunter-Gatherers and Early Herders in the Gobi Desert. J. World Prehist. *30*, 1–80.

Jeong, C., Ozga, A.T., Witonsky, D.B., Malmström, H., Edlund, H., Hofman, C.A., Hagan, R.W., Jakobsson, M., Lewis, C.M., Aldenderfer, M.S., et al. (2016). Long-term genetic stability and a high-altitude East Asian origin for the peoples of the high valleys of the Himalayan arc. Proc. Natl. Acad. Sci. USA *113*, 7485–7490.

Jeong, C., Wilkin, S., Amgalantugs, T., Bouwman, A.S., Taylor, W.T.T., Hagan, R.W., Bromage, S., Tsolmon, S., Trachsel, C., Grossmann, J., et al. (2018). Bronze Age population dynamics and the rise of dairy pastoralism on the eastern Eurasian steppe. Proc. Natl. Acad. Sci. USA *115*, E11248–E11255.

Jeong, C., Balanovsky, O., Lukianova, E., Kahbatkyzy, N., Flegontov, P., Zaporozhchenko, V., Immel, A., Wang, C.-C., Ixan, O., Khussainova, E., et al. (2019). The genetic history of admixture across inner Eurasia. Nat. Ecol. Evol. *3*, 966–976.

Jia, P.W.M., and Betts, A.V.G. (2010). A re-analysis of the Qiemu'erqieke (Shamirshak) cemeteries, Xinjiang, China. Journal of Indo-European Studies *38*, 275–318.

Jones, E.R., Gonzalez-Fortes, G., Connell, S., Siska, V., Eriksson, A., Martiniano, R., McLaughlin, R.L., Gallego Llorente, M., Cassidy, L.M., Gamba, C., et al. (2015). Upper Palaeolithic genomes reveal deep roots of modern Eurasians. Nat. Commun. *6*, 8912.

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics *29*, 1682–1684.

Jun, G., Wing, M.K., Abecasis, G.R., and Kang, H.M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Res. *25*, 918–925.

Keyser-Tracqui, C., Crubézy, E., and Ludes, B. (2003). Nuclear and mitochondrial DNA analysis of a 2,000-year-old necropolis in the Egyin Gol Valley of Mongolia. Am. J. Hum. Genet. *73*, 247–260.

Khatanbaatar, D. (2019). МОНГОЛЫН ЧУЛЛИЙН УЕИЙН ТУРУУЛГЛНЬ ЧАРУУЛСАН ОРШУУЛГАТ БУЛШНЫ СУДАЛГАА (Mongolyn khurliin ueiin turuulge ni kharuulsan orshuulgat bulshny sudalgaa) [The prone-positioned burials of Bronze Age in Mongolia]. PhD thesis (National University of Mongolia).

Kılınç, G.M., Omrak, A., Özer, F., Günther, T., Büyükkarakaya, A.M., Bıçakçı, E., Baird, D., Dönertaş, H.M., Ghalichi, A., Yaka, R., et al. (2016). The Demographic Development of the First Farmers in Anatolia. Curr. Biol. *26*, 2659–2666.

Kindstedt, P.S., and Ser-Od, T. (2019). Survival in a Climate of Change: The Origins and Evolution of Nomadic Dairying in Mongolia. Gastronomica *19*, 20–28.

Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics *15*, 356.

Kovalev, A. (2014). Earliest Europeans in the heart of Asia: the Chemurchek cultural phenomenon*Volume 1* (Saint Petersburg, Russia: Russian Academy of Sciences).

Kovalev, A. (2015). Earliest Europeans in the heart of Asia: the Chemurchek cultural phenomenon*2* (Saint Petersburg, Russia: Russian Academy of Sciences).

Kovalev, A. (2017). Munh-hairhanskaya kul'tura bronzovogo veka i ee svyazi s kul'turami Vostochnoi Sibiri [Munkh-Khairlkhan culture of Bronze Age and its connections with Neolithic-Bronze Age cultures of Eastern Siberia]. Actual Problems of Archaeology and Ethnology of Central Asia, pp. 58–66.

Kovalev, A., and Erdenebaatar, D. (2009). Discovery of new cultures of the Bronze Age in Mongolia according to the data obtained by the International Central Asian Archaeological Expedition. Current Archaeological Research in Mongolia *4*, 149–170.

Kovalev, A., Rukavishnikova, I.V., and Erdenebaatar, D. (2014). Olennye kamni – eto pamyatniki-kenotafy. Drevniye I Srednevekovye Izvayaniya Tsentralnoi Azii (Barnaul: Izd. Altai. Gos. Univ.), pp. 41–54.

Kovalev, A., Erdenebaatar, D., and Rukavishnikova, I.V. (2016). A Ritual complex with deer stones at Uushigiin Uvur, Mongolia: composition and construction stages. Archaeol. Ethnol. Anthropol. Eurasia *44*, 82–92.

Kradin, N.N. (2005). From Tribal Confederation to Empire: The Evolution of the Rouran Society. Acta Orientalia Academiae Scientiarum Hungaricae *58*, 149–169.

Kradin, N.N., and Ivliev, A.L. (2008). Deported nation: the fate of the Bohai people of Mongolia. Antiquity *82*, 438–445.

Krzewińska, M., Kılınç, G.M., Juras, A., Koptekin, D., Chyleński, M., Nikitin, A.G., Shcherbakov, N., Shuteleva, I., Leonova, T., Kraeva, L., et al. (2018). Ancient genomes suggest the eastern Pontic-Caspian steppe as the source of western Iron Age nomads. Sci Adv *4*, eaat4457.

Kubarev, V.D., and Shul'ga, P.I. (2007). Pazyrykskaya kul'tura (Barnaul: Altai State University).

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*, 409–413.

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., et al. (2016). Genomic insights into the origin of farming in the ancient Near East. Nature *536*, 419–424.

Lazaridis, I., Mittnik, A., Patterson, N., Mallick, S., Rohland, N., Pfrengle, S., Furtwängler, A., Peltzer, A., Posth, C., Vasilakis, A., et al. (2017). Genetic origins of the Minoans and Mycenaeans. Nature *548*, 214–218.

Lbova, L.V., Zhambaltarova, E.D., and Konev, V.P. (2008). Pogrebal'nye kompleksy Neolita - Rannego bronzovogo veka Zabaikal'ya. (Novosibirsk: Institut arkheologii i etnografii SO RAN).

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Gu, S., Han, Y., Xu, Z., Pakstis, A.J., Jin, L., Kidd, J.R., and Kidd, K.K. (2011). Diversification of the *ADH1B* gene during expansion of modern humans. Ann. Hum. Genet. *75*, 497–507.

Li, J., Zhang, Y., Zhao, Y., Chen, Y., Ochir, A., Sarenbilige, Zhu, H., and Zhou, H. (2018). The genome of an ancient Rouran individual reveals an important paternal lineage in the Donghu population. Am. J. Phys. Anthropol. *166*, 895–905.

Littleton, J., Floyd, B., Frohlich, B., Dickson, M., Amgalantögs, T., Karstens, S., and Pearlstein, K. (2012). Taphonomic analysis of Bronze Age burials in Mongolian khirigsuurs. J. Archaeol. Sci. *39*, 3361–3370.

Liu, W., Zhang, J., Wu, C., Cai, S., Huang, W., Chen, J., Xi, X., Liang, Z., Hou, Q., Zhou, B., et al. (2016). Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis. Sci. Rep. *6*, 34826.

Lkhagvasüren, K. (2007). Mongolyn Arkheologi (Chinges Khaany Ue). МОНГОЛЫН АРЧЕОЛОГИ (Ulaanbaatar: цИНГЗС ЧААНЫ УЕ).

Losey, R.J., Waters-Rist, A.L., Nomokonova, T., and Kharinskii, A.A. (2017). A Second Mortuary Hiatus on Lake Baikal in Siberia and the Arrival of Small-Scale Pastoralism. Sci. Rep. *7*, 2319.

Mackerras, C. (1972). The Uighur empire, according to the T'ang dynastic histories: a study in Sino-Uighur relations (Australian National University Press), pp. 744–840.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature *538*, 201–206.

Mann, A.E., Sabin, S., Ziesemer, K., Vågene, Å.J., Schroeder, H., Ozga, A.T., Sankaranarayanan, K., Hofman, C.A., Fellows Yates, J.A., Salazar-García, D.C., et al. (2018). Differential preservation of endogenous human and microbial DNA in dental calculus and dentin. Sci. Rep. *8*, 9822.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. Nature *528*, 499–503.

Mathieson, I., Alpaslan-Roodenberg, S., Posth, C., Szécsényi-Nagy, A., Rohland, N., Mallick, S., Olalde, I., Broomandkhoshbacht, N., Candilio, F., Cheronet, O., et al. (2018). The genomic history of southeastern Europe. Nature *555*, 197–203.

McColl, H., Racimo, F., Vinner, L., Demeter, F., Gakuhari, T., Moreno-Mayar, J.V., van Driem, G., Gram Wilken, U., Seguin-Orlando, A., de la Fuente Castro, C., et al. (2018). The prehistoric peopling of Southeast Asia. Science *361*, 88–92.

Miller, B.K. (2014). Xiongnu "Kings" and the Political Order of the Steppe Empire. J. Econ. Soc. Hist. Orient *57*, 1–43.

Miller, B.K. (2016). Xianbei Empire. In The Encyclopedia of Empire, N. Dalziel and J.M. MacKenzie, eds. (Oxford, UK: John Wiley & Sons, Ltd), pp. 1–3.

Miller, B.K., and Brosseder, U.B. (2017). Global dynamics in local processes of Iron Age Inner Asia. In The Routledge Handbook of Archaeology and Globalization, T. Hodos, A. Geurds, P. Lane, I. Lilley, M. Pitts, G. Shelach, M. Stark, and M.J. Versluys, eds. (Routledge), pp. 470–487.

Morgunova, N.L., and Khokhlova, O.S. (2013). Chronology and Periodization of the Pit-Grave Culture in the Region Between the Volga and Ural Rivers Based on Radiocarbon Dating and Paleopedological Research. Radiocarbon *55*, 1286–1296.

Murphy, E.M. (2003). Iron age archaeology and trauma from Aymyrlyg, south Siberia (British Archaeological Reports Limited).

Murphy, E.M., Schulting, R., Beer, N., Chistov, Y., Kasparov, A., and Pshenitsyna, M. (2013). Iron Age pastoral nomadism and agriculture in the eastern Eurasian steppe: implications from dental palaeopathology and stable carbon and nitrogen isotopes. J. Archaeol. Sci. *40*, 2547–2560.

Narasimhan, V.M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., Lazaridis, I., Nakatsuka, N., Olalde, I., Lipson, M., et al. (2019). The formation of human populations in South and Central Asia. Science *365*, eaat7487.

Nei Menggu zizhiqu wenwu kaogu yanjiusuo; International Institute for the Study of Nomadic Civilizations (2015). Mengguguo Bu'ergan sheng Daxinqileng sumu Zhanheshuo yizhi fajue jianbao [Brief report of excavations of remains at Zaan Khoshuu, Dashinchilin sum, Bulgan aimag, Mongolia]. Caoyuan Wenwu *2*, 8–31.

Nelson, A., Amartuvshin, C., and Honeychurch, W. (2009). A Gobi mortuary site through time: bioarchaeology at Baga Mongol, Baga Gazaryn Chuluu. Current Archaeological Research in Mongolia *4*, 565–578.

Novgorodova, E.A., Volkov, V.V., Korenevskij, S.N., and Mamonova, N.N. (1982). Ulangom: ein Graberfeld der skythischen Zeit aus der Mongolei (Wiesbaden: Harrassowitz).

Ochir, A., Danilov, S.V., Erdenebold, L., and Tserendorj, T. (2013). Ertnii nüüdelchdiin bunkhant bulshny maltlaga, sudalgaa (Ulaanbaatar: IISNC).

Ochir, A., Ankhbayar, B., and Tserenbyamba, K. (2016). Mongol, BHKhAU-yn khamtarsan arkheologiin kheeriin shinjilgeenii ajlyn ony tovch ür dün. In Mongolyn Arkheologi 2015, C. Amartüvshin, ed. (Ulaanbaatar: Mongolian Academy of Sciences), pp. 172–176.

Odbaatar, T. (2016). Elite Tombs of the Uyghur Period. In Ancient funeral monuments of Mongolia, G. Eregzen, ed. (Ulaanbaatar: Mongolian National Academy of Sciences), pp. 222–229.

Odbaatar, T., and Egiimaa, T. (2018). Syan'bi, Jujany üeiin tüükh, soyolyn sudalgaa (Ulaanbaatar: Mönkhiin Üseg), [Xianbei and Rouran period history and culture studies].

Ölziibayar, S., Ochir, B., and Urtnasan, E. (2019). Khünnügiin tüükh soyolyn sudalgaa (Ulaanbaatar: Mongolian Academy of Sciences), [Xiongnu history cultural studies].

Onon, U. (2005). The secret history of the Mongols: The life and times of Chinggis Khan (Routledge).

Parzinger, H. (2006). Die frühen Völker Eurasiens: vom Neolithikum bis zum Mittelalter. (C.H.Beck).

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. Genetics *192*, 1065–1093.

Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., and Nieselt, K. (2016). EAGER: efficient ancient genome reconstruction. Genome Biol. *17*, 60.

Poznik, G.D. (2016). Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. bioRxiv. https://doi.org/10.1101/088716.

Raghavan, M., Skoglund, P., Graf, K.E., Metspalu, M., Albrechtsen, A., Moltke, I., Rasmussen, S., Stafford, T.W., Jr., Orlando, L., Metspalu, E., et al. (2014). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature *505*, 87–91.

Raghavan, M., Steinrücken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., Albrechtsen, A., Valdiosera, C., Ávila-Arcos, M.C., Malaspinas, A.-S., et al. (2015). POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science *349*, aab3884.

Ramsey, C.B. (2017). Methods for Summarizing Radiocarbon Datasets. Radiocarbon *59*, 1809–1833.

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J.S., Albrechtsen, A., Moltke, I., Metspalu, M., Metspalu, E., Kivisild, T., Gupta, R., et al. (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature *463*, 757–762.

Rasmussen, M., Anzick, S.L., Waters, M.R., Skoglund, P., DeGiorgio, M., Stafford, T.W., Jr., Rasmussen, S., Moltke, I., Albrechtsen, A., Doyle, S.M., et al. (2014). The genome of a Late Pleistocene human from a Clovis burial site in western Montana. Nature *506*, 225–229.

Rasmussen, M., Sikora, M., Albrechtsen, A., Korneliussen, T.S., Moreno-Mayar, J.V., Poznik, G.D., Zollikofer, C.P.E., de León, M.P., Allentoft, M.E., Moltke, I., et al. (2015). The ancestry and affiliations of Kennewick Man. Nature *523*, 455–458.

Reimer, P.J., Bard, E., Bayliss, A., Warren Beck, J., Blackwell, P.G., Ramsey, C.B., Buck, C.E., Cheng, H., Lawrence Edwards, R., Friedrich, M., et al. (2013). IntCal13 and Marine13 Radiocarbon Age Calibration Curves 0–50,000 Years cal BP. Radiocarbon *55*, 1869–1887.

Renaud, G., Slon, V., Duggan, A.T., and Kelso, J. (2015). Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. Genome Biol. *16*, 224.

Rogers, L. (2016). Understanding ancient human population genetics of the eastern Eurasian steppe through mitochondrial DNA analysis: Central Mongolian samples from the Neolithic, Bronze Age, Iron Age and Mongol Empire periods. PhD thesis (Indiana University).

Rogers, L.L., Honeychurch, W., Amartuvshin, C., and Kaestle, F.A. (2020). U5a1 Mitochondrial DNA Haplotype Identified in Eneolithic Skeleton from Shatar Chuluu, Mongolia. Human Biology *91*, 213–223.

Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., and Reich, D. (2015). Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20130624.

Rudenko, S.I. (1970). Frozen Tombs of Siberia: The Pazyryk Burials of Iron Age Horsemen, M.W. Thompson, ed. (Berkeley, Los Angeles: University of California Press), trans.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al.; International HapMap Consortium (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.

Samashev, Z. (2011). Berel (Almaty: TBU).

Savinov, D.G. (2002). Rannie kochevniki Verkhnego Yeniseya (St. Petersburg: St. Petersburg State University), [Early Nomads of Upper Yenisei].

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., and Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. PLoS ONE *7*, e34131.

Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res. Notes *9*, 88.

Sikora, M., Pitulko, V.V., Sousa, V.C., Allentoft, M.E., Vinner, L., Rasmussen, S., Margaryan, A., de Barros Damgaard, P., de la Fuente, C., Renaud, G., et al. (2019). The population history of northeastern Siberia since the Pleistocene. Nature *570*, 182–188.

Siska, V., Jones, E.R., Jeon, S., Bhak, Y., Kim, H.-M., Cho, Y.S., Kim, H., Lee, K., Veselovskaya, E., Balueva, T., et al. (2017). Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago. Sci. Adv. *3*, e1601877.

Taylor, W.T.T., Bayarsaikhan, J., and Tuvshinjargal, T. (2015). Equine cranial morphology and the identification of riding and chariotry in late Bronze Age Mongolia. Antiquity *89*, 854–871.

Taylor, W.T.T., Bayarsaikhan, J., Tuvshinjargal, T., Bender, S., Tromp, M., Clark, J., Lowry, K.B., Houle, J.L., Staszewski, D., Whitworth, J., and Fitzhugh, W. (2018). Origins of equine dentistry. Proceedings of the National Academy of Sciences *115*, E6707–E6715.

Taylor, W.T.T., Jargalan, B., Lowry, K.B., Clark, J., Tuvshinjargal, T., and Bayarsaikhan, J. (2017). A Bayesian chronology for early domestic horse use in the Eastern Steppe. J. Archaeol. Sci. *81*, 49–58.

Taylor, W., Wilkin, S., Wright, J., Dee, M., Erdene, M., Clark, J., Tuvshinjargal, T., Bayarsaikhan, J., Fitzhugh, W., and Boivin, N. (2019). Radiocarbon dating and cultural dynamics across Mongolia's early pastoral transition. PLoS ONE *14*, e0224241.

Törbat, T. (2016). Sagsai-shaped graves. In Ancient Funeral Monuments of Mongolia, D. Eregzen, ed. (Institute of History and Archaeology, Mongolian Academy of Sciences), pp. 58–71.

Törbat, T., Giscard, P.H., and Batsükh, D. (2009). First excavation of Pazyryk kurgans in Mongolian Altai (Bonn: Current Archaeological Research in Mongolia. Universität Bonn), pp. 221–231.

Tseveendorj, D. (1980). Chandmanii Soyol (Ulaanbaatar: Mongolian Academy of Sciences).

Tseveendorj, D. (2007). Chandmanii soyol (Nemen zasvarlasan khoyor dakh' khevlel) (Ulaanbaatar: Mongolian Academy of Sciences).

Tsybiktarov, A.D. (1998). Kultura plitochnyih mogil Mongolii i Zabaikalya [Slab Grave Culture of Mongolia and Transbaikalia]. (BSU).

Tsybiktarov, A. (2003). Central Asia in the Bronze and Early Iron Ages: Problems of ethnocultural history of Mongolia and the Southern Transbaikal region in the middle second to early first millennia BC. Archaeol. Ethnol. Anthropol. Eurasia *13*, 80–97.

Unterländer, M., Palstra, F., Lazaridis, I., Pilipenko, A., Hofmanová, Z., Groß, M., Sell, C., Blöcher, J., Kirsanow, K., Rohland, N., et al. (2017). Ancestry and demography and descendants of Iron Age nomads of the Eurasian Steppe. Nat. Commun. *8*, 14615.

Vadetskaya, E., Polyakov, A., and Stepanova, N. (2014). Svod pamyatnikov afanas'evskoi kul'tury (Barnaul: AZBUKA), [Corpus of Afanasievo culture sites].

Ventresca Miller, A.R., and Makarewicz, C.A. (2019). Intensification in pastoralist cereal use coincides with the expansion of trans-regional networks in the Eurasian Steppe. Sci. Rep. *9*, 8363.

Volkov, V. (2002). Olennye kamni Mongolii (Moscow: Nauchnyi Mir).

Wang, C.-C., Reinhold, S., Kalmykov, A., Wissgott, A., Brandt, G., Jeong, C., Cheronet, O., Ferry, M., Harney, E., Keating, D., et al. (2019). Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. Nat. Commun. *10*, 590.

Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.-J., Kronenberg, F., Salas, A., and Schönherr, S. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. *44* (W1), W58–W63.

Wilkin, S., Ventresca Miller, A., Taylor, W.T.T., Miller, B.K., Hagan, R.W., Bleasdale, M., Scott, A., Gankhuyg, S., Ramsøe, A., Uliziibayar, S., et al. (2020a). Dairy pastoralism sustained eastern Eurasian steppe populations for 5,000 years. Nat. Ecol. Evol. *4*, 346–355.

Wilkin, S., Ventresca Miller, A., Miller, B.K., Spengler, R.N., 3rd, Taylor, W.T.T., Fernandes, R., Hagan, R.W., Bleasdale, M., Zech, J., Ulziibayar, S., et al. (2020b). Economic diversification supported the growth of Mongolia's nomadic empires. Sci. Rep. *10*, 3916.

Wright, J.S.C. (2012). Temporal perspectives on the monumental constellations of Inner Asia. "As Time Goes By?" Monumentality, Landscapes and the Temporal Perspective (Rudolf Habelt).

Yang, M.A., Gao, X., Theunert, C., Tong, H., Aximu-Petri, A., Nickel, B., Slatkin, M., Meyer, M., Pääbo, S., Kelso, J., and Fu, Q. (2017). 40,000-Year-Old Individual from Asia Provides Insight into Early Population Structure in Eurasia. Curr. Biol. *27*, 3202–3208.e9.

Zazzo, A., Lepetz, S., Magail, J., and Gantulga, J.-O. (2019). High-precision dating of ceremonial activity around a large ritual complex in Late Bronze Age Mongolia. Antiquity *93*, 80–98.

Zerjal, T., Xue, Y., Bertorelle, G., Wells, R.S., Bao, W., Zhu, S., Qamar, R., Ayub, Q., Mohyuddin, A., Fu, S., et al. (2003). The genetic legacy of the Mongols. Am. J. Hum. Genet. *72*, 717–721.

Zwyns, N., Paine, C.H., Tsedendorj, B., Talamo, S., Fitzsimmons, K.E., Gantumur, A., Guunii, L., Davakhuu, O., Flas, D., Dogandžić, T., et al. (2019). The Northern Route for Human dispersal in Central and Northeast Asia: New evidence from the site of Tolbor-16, Mongolia. Sci. Rep. *9*, 11759.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Human archaeological skeletal material	This study	ARG001(AT-765)
Human archaeological skeletal material	This study	ARG002(AT-764)
Human archaeological skeletal material	This study	ARG003(AT-763)
Human archaeological skeletal material	This study	AST001(AT-841)
Human archaeological skeletal material	This study	ATS001(AT-459)
Human archaeological skeletal material	This study	BAM001(AT-752)
Human archaeological skeletal material	This study	BAU001(AT-409)
Human archaeological skeletal material	This study	BAY001(AT-304)
Human archaeological skeletal material	This study	BAZ001(AT-846)
Human archaeological skeletal material	This study	BER002(AT-905)
Human archaeological skeletal material	This study	BIL001(AT-340)
Human archaeological skeletal material	This study	BOR001(AT-707)
Human archaeological skeletal material	This study	BRG001(AT-650)
Human archaeological skeletal material	This study	BRG002(AT-651)
Human archaeological skeletal material	This study	BRG004(AT-655)
Human archaeological skeletal material	This study	BRG005(AT-653)
Human archaeological skeletal material	This study	BRL001(AT-296)
Human archaeological skeletal material	This study	BRL002(AT-294)
Human archaeological skeletal material	This study	BRU001(AT-154)
Human archaeological skeletal material	This study	BTO001(AT-435)
Human archaeological skeletal material	This study	BUL001(AT-923)
Human archaeological skeletal material	This study	BUL002(AT-922)
Human archaeological skeletal material	This study	BUR001(AT-589)
Human archaeological skeletal material	This study	BUR002(AT-536)
Human archaeological skeletal material	This study	BUR003(AT-535)
Human archaeological skeletal material	This study	BUR004(AT-537)
Human archaeological skeletal material	This study	CHD001(AT-173)
Human archaeological skeletal material	This study	CHN001(AT-121)
Human archaeological skeletal material	This study	CHN003(AT-141)
Human archaeological skeletal material	This study	CHN004(AT-105)
Human archaeological skeletal material	This study	CHN006(AT-109)
Human archaeological skeletal material	This study	CHN007(AT-128)
Human archaeological skeletal material	This study	CHN008(AT-138)
Human archaeological skeletal material	This study	CHN010(AT-119)
Human archaeological skeletal material	This study	CHN012(AT-98)
Human archaeological skeletal material	This study	CHN014(AT-125)
Human archaeological skeletal material	This study	CHN015(AT-115)
Human archaeological skeletal material	This study	CHN016(AT-208)
Human archaeological skeletal material	This study	DAR001(AT-766)
Human archaeological skeletal material	This study	DAR002(AT-767)
Human archaeological skeletal material	This study	DAS001(AT-391)
Human archaeological skeletal material	This study	DEE001(AT-389)
Human archaeological skeletal material	This study	DEK001/SHR001(AT-755)

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Human archaeological skeletal material	This study	DEL001(AT-530)
Human archaeological skeletal material	This study	DOL001(AT-370)
Human archaeological skeletal material	This study	DUU001(AT-605)
Human archaeological skeletal material	This study	DUU002(AT-407)
Human archaeological skeletal material	This study	EME002(AT-708)
Human archaeological skeletal material	This study	ERD001(AT-831)
Human archaeological skeletal material	This study	ERM001/ERM002/ ERM003(DA-KG-1909-001)
Human archaeological skeletal material	This study	FNO001(2008, pogrebenie 3)
Human archaeological skeletal material	This study	FNO003(2008, pogrebenie 4, skeleton 2)
Human archaeological skeletal material	This study	FNO006(2007, pogrebenie 1, formerly pogrebenie 18, main individual)
Human archaeological skeletal material	This study	FNO007(1996, pogrebenie 11, kostyak 2)
Human archaeological skeletal material	This study	GAN002(AT-835)
Human archaeological skeletal material	This study	GTO001(AT-624)
Human archaeological skeletal material	This study	GUN002(AT-780)
Human archaeological skeletal material	This study	HUD001(AT-290)
Human archaeological skeletal material	This study	IAG001(AT-590B)
Human archaeological skeletal material	This study	IKU001(AT-772)
Human archaeological skeletal material	This study	IMA001(2006 Mogila 76)
Human archaeological skeletal material	This study	IMA002(2005 Mogila 75)
Human archaeological skeletal material	This study	IMA003(2005 Mogila 73)
Human archaeological skeletal material	This study	IMA004(2003 Mogila 70)
Human archaeological skeletal material	This study	IMA005(2007 Mogila 78)
Human archaeological skeletal material	This study	IMA006(2007 Mogila 77)
Human archaeological skeletal material	This study	IMA007(2007 Mogila 79)
Human archaeological skeletal material	This study	IMA008(2004 Mogila 71)
Human archaeological skeletal material	This study	JAA001(AT-910)
Human archaeological skeletal material	This study	JAG001(AT-878)
Human archaeological skeletal material	This study	KGK001(AT-900)
Human archaeological skeletal material	This study	KHI001(AT-398)
Human archaeological skeletal material	This study	KHL001(AT-363)
Human archaeological skeletal material	This study	KHN001/KHN002 (AT-758; AT-759)
Human archaeological skeletal material	This study	KHO001(AT-354)
Human archaeological skeletal material	This study	KHO006(AT-361B)
Human archaeological skeletal material	This study	KHO007(AT-361A)
Human archaeological skeletal material	This study	KHU001(AT-861)
Human archaeological skeletal material	This study	KHV002(AT-811)
Human archaeological skeletal material	This study	KNN001(AT-754)
Human archaeological skeletal material	This study	KNU001(AT-352)
Human archaeological skeletal material	This study	KRN001(AT-643)
Human archaeological skeletal material	This study	KRN002(AT-644)
Human archaeological skeletal material	This study	KUM001(AT-628)
Human archaeological skeletal material	This study	KUR001(AT-635)

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Human archaeological skeletal material	This study	MIT001(AT-975)
Human archaeological skeletal material	This study	MRI001(AT-800)
Human archaeological skeletal material	This study	NAI001(AT-149)
Human archaeological skeletal material	This study	NAI002/NAI003(AT-152)
Human archaeological skeletal material	This study	NOM001(AT-917)
Human archaeological skeletal material	This study	NRC001(AT-393)
Human archaeological skeletal material	This study	OLN001.A(AT-871)
Human archaeological skeletal material	This study	OLN001.B(AT-871)
Human archaeological skeletal material	This study	OLN002(AT-891)
Human archaeological skeletal material	This study	OLN003(AT-892)
Human archaeological skeletal material	This study	OLN004(AT-969)
Human archaeological skeletal material	This study	OLN005(AT-973)
Human archaeological skeletal material	This study	OLN007(AT-972)
Human archaeological skeletal material	This study	OLN008(AT-873)
Human archaeological skeletal material	This study	OLN009(AT-896)
Human archaeological skeletal material	This study	OLN010(AT-893)
Human archaeological skeletal material	This study	OLN011(AT-897)
Human archaeological skeletal material	This study	OLN012(AT-894)
Human archaeological skeletal material	This study	PTO001 (Plitochnaya Mogila 4)
Human archaeological skeletal material	This study	RAH001(AT-532)
Human archaeological skeletal material	This study	SAN001(AT-575)
Human archaeological skeletal material	This study	SBG001(AT-960)
Human archaeological skeletal material	This study	SHA001(AT-594)
Human archaeological skeletal material	This study	SHG001(AT-701)
Human archaeological skeletal material	This study	SHG002(AT-699)
Human archaeological skeletal material	This study	SHG003(AT-703)
Human archaeological skeletal material	This study	SHT001(AT-26)
Human archaeological skeletal material	This study	SHT002(AT-25)
Human archaeological skeletal material	This study	SHU001(AT-233)
Human archaeological skeletal material	This study	SHU002(AT-232B)
Human archaeological skeletal material	This study	SKT001(CA-4-1)
Human archaeological skeletal material	This study	SKT002(CA-19)
Human archaeological skeletal material	This study	SKT003(CA-13-1)
Human archaeological skeletal material	This study	SKT004(CA-24)
Human archaeological skeletal material	This study	SKT005(CA-8)
Human archaeological skeletal material	This study	SKT006(CA-17)
Human archaeological skeletal material	This study	SKT007(CA-3-1)
Human archaeological skeletal material	This study	SKT008(CA-28)
Human archaeological skeletal material	This study	SKT009(CA-9-1)
Human archaeological skeletal material	This study	SKT010(CA-7)
Human archaeological skeletal material	This study	SKT012(CA-29)
Human archaeological skeletal material	This study	SOL001(AT-274)
Human archaeological skeletal material	This study	SON001(AT-150)
Human archaeological skeletal material	This study	SOU001(AT-501)
Human archaeological skeletal material	This study	TAH002(AT-360)
Human archaeological skeletal material	This study	TAK001(AT-401A)
Human archaeological skeletal material	This study	TAV001(AT-625/688)

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Human archaeological skeletal material	This study	TAV005(AT-670/695)
Human archaeological skeletal material	This study	TAV006(AT-623)
Human archaeological skeletal material	This study	TAV011(AT-671/687)
Human archaeological skeletal material	This study	TEV002(AT-33)
Human archaeological skeletal material	This study	TEV003(AT-145)
Human archaeological skeletal material	This study	TMI001(AT-751)
Human archaeological skeletal material	This study	TSA001(AT-784)
Human archaeological skeletal material	This study	TSA002(AT-816)
Human archaeological skeletal material	This study	TSA003(AT-783)
Human archaeological skeletal material	This study	TSA004(AT-782)
Human archaeological skeletal material	This study	TSA005(AT-815)
Human archaeological skeletal material	This study	TSA006(AT-814)
Human archaeological skeletal material	This study	TSA007(AT-786)
Human archaeological skeletal material	This study	TSB001(AT-804)
Human archaeological skeletal material	This study	TSI001(AT-802)
Human archaeological skeletal material	This study	TUH001(AT-543)
Human archaeological skeletal material	This study	TUH002(AT-542)
Human archaeological skeletal material	This study	TUK001/TAV008 (AT-729;AT-728)
Human archaeological skeletal material	This study	TUK002(AT-757)
Human archaeological skeletal material	This study	TUK003(AT-684)
Human archaeological skeletal material	This study	TUM001(AT-913)
Human archaeological skeletal material	This study	UAA001(AT-614)
Human archaeological skeletal material	This study	UGO001(AT-588)
Human archaeological skeletal material	This study	UGO002(AT-581)
Human archaeological skeletal material	This study	UGU001(AT-749)
Human archaeological skeletal material	This study	UGU002(AT-549)
Human archaeological skeletal material	This study	UGU003(AT-570)
Human archaeological skeletal material	This study	UGU004(AT-805)
Human archaeological skeletal material	This study	UGU005(AT-747)
Human archaeological skeletal material	This study	UGU006(AT-692)
Human archaeological skeletal material	This study	UGU010(AT-690)
Human archaeological skeletal material	This study	UGU011(AT-748)
Human archaeological skeletal material	This study	ULA001(AT-840)
Human archaeological skeletal material	This study	ULI001(AT-676)
Human archaeological skeletal material	This study	ULI002(AT-675)
Human archaeological skeletal material	This study	ULI003(AT-680)
Human archaeological skeletal material	This study	ULN001(AT-823)
Human archaeological skeletal material	This study	ULN002(AT-920)
Human archaeological skeletal material	This study	ULN003(AT-921)
Human archaeological skeletal material	This study	ULN004(AT-885)
Human archaeological skeletal material	This study	ULN005(AT-769)
Human archaeological skeletal material	This study	ULN006(AT-962)
Human archaeological skeletal material	This study	ULN007(AT-883)
Human archaeological skeletal material	This study	ULN009(AT-884)
Human archaeological skeletal material	This study	ULN010(AT-964)
Human archaeological skeletal material	This study	ULN011(AT-882)
Human archaeological skeletal material	This study	ULN015(AT-824)

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Human archaeological skeletal material	This study	ULZ001(AT-674)
Human archaeological skeletal material	This study	UUS001(AT-613)
Human archaeological skeletal material	This study	UUS002(AT-610)
Human archaeological skeletal material	This study	UVG001(AT-338)
Human archaeological skeletal material	This study	YAG001(AT-590A)
Human archaeological skeletal material	This study	YUR001(AT-649)
Human archaeological skeletal material	This study	ZAA001(AT-954)
Human archaeological skeletal material	This study	ZAA002(AT-957)
Human archaeological skeletal material	This study	ZAA003(AT-953)
Human archaeological skeletal material	This study	ZAA004(AT-959)
Human archaeological skeletal material	This study	ZAA005(AT-956)
Human archaeological skeletal material	This study	ZAA007(AT-958)
Human archaeological skeletal material	This study	ZAM001(AT-390)
Human archaeological skeletal material	This study	ZAM002(AT-711)
Human archaeological skeletal material	This study	ZAR002(AT-271)
Human archaeological skeletal material	This study	ZAY001(AT-768)
Chemicals, Peptides, and Recombinant Proteins		
USER™ Enzyme, recombinant	NEB	M5508
Critical Commercial Assays		
HiSeq® 3000/4000 SR Cluster Kit	Illumina	PE-410-1001
HiSeq® 3000/4000 PE Cluster Kit	Illumina	GD-410-1001
HiSeq® 3000/4000 SBS Kit (50 cycles)	Illumina	FC-410-1001
HiSeq® 3000/4000 SBS Kit (150 cycles)	Illumina	FC-410-1002
Deposited Data		
Raw and analyzed data	This study	ENA: PRJEB35748
Haploid genotype data for 1240K panel (Edmond Data Repository of the Max Planck Society)	This study	https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo
Software and Algorithms		
EAGER v1.92.55	(Peltzer et al., 2016)	https://github.com/apeltzer/EAGER-GUI
AdapterRemoval v2.2.20	(Schubert et al., 2016)	https://github.com/MikkelSchubert/adapterremoval
BWA v0.7.12	(Li and Durbin, 2009)	http://bio-bwa.sourceforge.net
dedup v0.12.2	(Peltzer et al., 2016)	https://github.com/apeltzer/DeDup
bamUtils v.1.0.13	(Jun et al., 2015)	https://github.com/statgen/bamUtil
samtools mpileup	(Li and Durbin, 2009)	http://www.htslib.org/doc/samtools.html
pilupCaller v1.2.2	(https://github.com/stschiff/sequenceTools)	https://github.com/stschiff/sequenceTools
mapDamage v2.0.6	(Jónsson et al., 2013)	https://github.com/MikkelSchubert/mapDamage
Schmutzi	(Renaud et al., 2015)	https://github.com/grenaud/schmutzi
circularmapper v1.1	(Peltzer et al., 2016)	https://github.com/apeltzer/CircularMapper
ANGSD v0.910	(Korneliussen et al., 2014)	http://www.popgen.dk/angsd/index.php/ANGSD

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
HaploGrep 2 v2.1.19	(Weissensteiner et al., 2016)	https://haplogrep.i-med.ac.at/category/haplogrep2/
yHaplo	(Poznik, 2016)	https://github.com/alexhbnr/yhaplo
Eigensoft v7.2.1	(Patterson et al., 2006)	https://github.com/DReichLab/EIG
DATES	(Narasimhan et al., 2019)	https://github.com/priyamoorjani/DATES
admixtools v5.1	(Patterson et al., 2012)	https://github.com/DReichLab/AdmixTools

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Christina Warinner (warinner@fas.harvard.edu).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The accession number for all newly reported sequencing data reported in this paper are available from the European Nucleotide Archive: PRJEB35748. 1240K genotype data are available on the Edmond Max Planck Data Repository under the link: https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Here we present new genome-wide data for 213 ancient individuals from Mongolia and 13 individuals from Buryatia, Russia, which we analyze together with 21 previously published ancient Mongolian individuals (Jeong et al., 2018), for a total of 247 individuals. Human remains analyzed in this study were reviewed and approved by the Mongolian Ministry of Culture and the Mongolian Ministry of Education, Culture, Science, and Sport under reference numbers A0122772 MN DE 0 8124, A0109258 MN DE 7 643, and A0117901 MN DE 9 4314, and declaration number 12-2091008-20E00225. All new Mongolian individuals, except ERM, were sampled from the physical anthropology collections at the National University of Mongolia and the Institute for Archaeology and Ethnology in Ulaanbaatar, Mongolia. ERM001/002/003 was provided by Jan Bemmann. Russian samples were collected from the Institute for Mongolian, Buddhist, and Tibetan Research as well as the Buryat Scientific Center, Russian Academy of Sciences (RAS).

Together, this ancient Eastern Steppe dataset of 247 individuals originates from 89 archaeological sites (Figure 1; Figure S1A; Table S1A) and spans approximately 6,000 years of time (Tables S1A, S1B, and S2C). High quality genetic data was successfully generated for 214 individuals and was used for population genetic analysis (Table S2A). Subsistence information inferred from proteomic analysis of dental calculus has been recently published for a subset of these individuals (n = 32; Wilkin et al., 2020a), and stable isotope analysis of bone collagen and enamel (n = 137) is also in progress (Wilkin et al., 2020b); together, these data allow direct comparison between the biological ancestry of specific archaeological cultures and their diets, particularly with respect to their dairy and millet consumption. Below, we provide an overview of the geography and ecology of the archaeological sites in this study, as well as their temporal and cultural context.

Geography and ecology of Mongolia

Mongolia is located in Inner Asia between Russia and China, and it encompasses most of the Eurasian Eastern Steppe (Figure 1; Figure S1A). Mongolia has 21 aimags (provinces) and can be divided into ten geographic regions (Figure S1B) with distinct ecological (Figure S1C) and cultural features (Taylor et al., 2019). For example, far north Mongolia borders Siberia and includes both high mountain and mountain-taiga ecological zones, and it is the only aimag where reindeer pastoralism is practiced. North Mongolia is dominated by forest-steppe, but also contains mountain-taiga and steppe zones; cattle and yak pastoralism is particularly productive here, and Bulgan province is renowned for its horse pastoralism. The Altai region represents an extension of the Altai mountains from Russia into Mongolia and consists of a patchwork of environments including high mountains, valleys, and lakes, and ranging from forest steppe to desert steppe as the region stretches from north to south; pastoral economy in the Altai is mixed and differs

by local environmental conditions. South Mongolia is dominated by the Gobi Desert, and it borders central and southeast Mongolia, which are largely characterized by desert-steppe; camel pastoralism is found throughout these regions. East Mongolia is a large expansive steppe zone that stretches to northeastern China. Today, mining is important in eastern Mongolia, as well as cattle, sheep, goat, and horse pastoralism.

Overview of Mongolian archaeology

Mongolian prehistory extends back more than 40,000 years, with documented sites ranging from the Upper Paleolithic to the present day. During nearly all of this time, lifeways in Mongolia have been nomadic, either supported by hunting, fishing and gathering or by pastoralism. The short-term and ephemeral nature of nomadic camp sites makes them difficult to identify on the landscape, and wind deflation has further reduced the visibility and preservation of many domestic sites. Only during the Bronze Age, with the sudden appearance of stone mounds and other burial features, do sites become more conspicuous and the archaeology better attested. As such, knowledge of Mongolian prehistory is strongly biased toward the past five millennia. The archaeology of Mongolia can be divided into 7 main periods: (1) pre-Bronze Age, prior to 3500 BCE; (2) Early Bronze Age, 3500-1900 BCE; (3) Middle/Late Bronze Age, 1900-900 BCE; (4) Early Iron Age, 900-300 BCE; (5) Xiongnu, 200 BCE to 100 CE; (6) Early Medieval, 100-850 CE, and (7) Late Medieval, 850-1650 CE. A brief summary of each period, as well as details for the sites included in this study, are provided below, in Figure S1, and in Table S1.

Pre-Bronze Age (prior to 3500 BCE)

The early archaeological record of Mongolia is poorly understood, particularly with respect to human remains and burials. While occasional finds provide direct evidence of anatomically modern humans in Mongolia as far back as the Early Upper Paleolithic (Devièse et al., 2019; Zwyns et al., 2019), only a small handful of intentionally buried skeletons have been recovered prior to the end of the 4th millennium BCE. Although early and middle Holocene-era (10,000-3500 BCE) features and burials have been referred to as "Mesolithic," "Neolithic," or "Eneolithic" (Hanks, 2010), there is no direct evidence for domestic animals or a food-producing economy at any of these localities, although pottery was in wide use by the mid-Holocene (Janz et al., 2017). Pre-Bronze Age burials in eastern Mongolia are characterized by an absence of surficial construction features, while in northern Mongolia pre-Bronze burials typically consist of small stone cairns. The burial goods of this period include artifacts made from stone, mother-of-pearl, and animal bones, such as deer and marmot (Eregzen, 2016). Two individuals in this study date to this pre-Bronze Age period in Mongolia. The first, dating to ca. 4600 BCE, is from Kherlengiin Ereg (SOU), located on the south bank of the Kherlen River near Choibalsan city at the extreme eastern end of Mongolia in Dornod province (Dorj, 1969). It was found in a disturbed context, and the original burial position could not be reconstructed. However, other burials from this period and region are typically crouched (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). The second, dating to ca. 3700 BCE, is from Erdenemandal (ERM) in the Arkhangai province of north Mongolia (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). It lacked stone construction features and consisted of a simple crouched pit burial beneath a shallow earthen mound. The burial was recovered at a depth of 2.3 m and the grave mound appears to be part of a larger cemetery.

In addition to these two pre-Bronze burials from Mongolia, we also analyzed four pre-Bronze individuals from the site of Fofonovo (FNO) near Lake Baikal in Buryatia, Russia (Lbova et al., 2008). All were buried in simple pit burials partially flexed (only their legs bent) as individuals, or sometimes several persons together. People at Fofonovo, like many others around Lake Baikal, were interred with many burial goods, including an array of bone and stone beads, neck pieces, chipped stone blades and points, bone harpoons, and pottery. Among these items were fragments or worked ornaments from wild boar, sable, and hawk.

Although pre-Bronze Age material from Mongolia is sparse, recent excavations in neighboring regions provide important context. In southern Russia, excavations by the Baikal Archaeological Project (BAP) at three sites (Lokomotiv, Shamanka II, and Ust'-Ida I) have enabled characterization of the Lake Baikal Neolithic Kitoi (5200-4200 BCE) and Isakovo (4000-3000 BCE) mortuary traditions, including genome sequencing of 14 of these hunter-gatherers (Damgaard et al., 2018a) (Baikal_EN; also characterized in later studies as East Siberian Hunter Gatherers, ESHG) (Narasimhan et al., 2019). The genomes of six hunter-gatherers dating to ca. 5700 BCE are also available from the site of Devil's Gate (Sikora et al., 2019; Siska et al., 2017), a Neolithic cave site on the border between Russia and Korea. These individuals, separated by 2500 km, share a similar ancestry to each other and to modern Tungusic speakers in the lower Amur Basin, who we refer to as Ancient Northeast Asians (ANA).

Early Bronze Age (ca. 3500-1900 BCE)

Two major cultural phenomena associated with monumental mortuary architecture have been described in Mongolia during the Early Bronze Age (EBA): Afanasievo and Chemurchek. Both exhibit features linking them to ruminant pastoralism and to cultures further west.

Afanasievo (3150-2750 BCE). Beginning ca. 3150 BCE and persisting until ca. 2750 BCE (Taylor et al., 2019), although perhaps as late as 2600 BCE, stone burials belonging to the Afanasievo culture type have been recovered from the Khangai Mountains in central Mongolia and the Altai Mountains of western Mongolia. These burials contain the earliest direct evidence for domestic livestock (sheep/goat and cattle) in Mongolia. Afanasievo burials in Mongolia typically consist of circular flat stones bounded by upright stones, which overlay an internal burial pit containing an extended individual with flexed legs. Such burials are similar to Afanasievo kurgans in the Russian Altai (Vadetskaya et al., 2014), which contain burials of supine individuals with flexed legs and heads typically facing east. The burial mounds at Khuurai Gobi 1 and Ulaankhus (Bayan-Ulgii province, western Mongolia; not sampled in this study) exhibit typical Afanasievo architectural features (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo).

In addition to domestic animal remains, Afanasievo burial mounds contain egg-shaped pottery vessels, and sometimes include metal artifacts (from copper, gold, and silver) and apparent deconstructed cart objects (Kovalev and Erdenebaatar, 2009). Recent analysis of proteins in human dental calculus from Afanasievo burials directly demonstrates the utilization of ruminant dairy products and the presence of domestic animals in the Afanasievo economy (Wilkin et al., 2020a).

We analyzed individuals from one Afanasievo site in this study: Shatar Chuluu (SHT). Located in Byankhongor province on the south slope of the Khangai mountains, Shatatar Chuulu is the easternmost known Afanasievo cemetery in Eurasia. Three of the site's burial mounds have been excavated, and each consisted of a flat platform of round stones bounded by large boulders. The burials were arranged in pits beneath the mounds, and the bodies were laid out in a supine position with flexed knees and heads facing to the west. Despite the fact that few burial goods were found, the overall architectural design of the mounds combined with isolated fragments of typical Afanasievo vessels make it possible to attribute these mounds to the Afanasievo archaeological culture.

Chemurchek (2750-1900 BCE). The Chemurchek archaeological culture (also called Hemtseg, Qiemu'erqieke, Shamirshak), spans the period between 2750 BCE-1900 BCE (Taylor et al., 2019). These features are found in western Mongolia and adjoining regions of bordering countries, including the Dzungarian Basin of Xinjiang and eastern Kazakhstan (Honeychurch, 2017; Kovalev and Erdenebaatar, 2009). Chemurchek mortuary architecture is characterized by collective burials in large stone cists surrounded by stone and earthen cairns overlapping one another or by large rectangular stone fences up to 50 m in length. The Chemurcheck burial at Kheviin Am (Khovd province, western Mongolia; not sampled in this study) exhibits typical Chemurchek features (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). Similar to Afanasievo burials, individuals found in Chemurchek tombs are laid out in a supine position with flexed legs. Adjacent to many Chemurchek burial features along the eastern side are anthropomorphic standing stones, sometimes depicted holding a shepherd's crook (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). Inside the burials, artifacts such as stone bowls, bone tools, ceramics, and sometimes metal jewelry, have been recovered. Occasionally, non-funerary ritual structures, such as fences containing earthen pits with charcoal and animal remains or large stone fences depicting petroglyphs, are also attributed to this culture (Kovalev, 2014). Recent analysis of proteins in human dental calculus from these features confirmed the utilization of ruminant dairy products and the presence of domestic animals in the Chemurchek economy (Wilkin et al., 2020a), although available radiocarbon chronology appears to preclude a meaningful exploitation of domestic horses (Taylor et al., 2019). We analyzed individuals from three Chemurchek sites in this study: Yagshiin Khuduu (IAG/YAG), Khundii Gobi (KUM), and Khuurai Gobi 2 (KUR). Yagshiin Khuduu is located in the southern Mongolian Altai, while Khundii Gobi and Khuurai Gobi 2 are located in the northern Mongolian Altai. Whereas Yagshiin Khuduu represents a typical Chemurchek burial within a stone cist, the two northern Chemurchek mortuary sites consist of burials within rectangular mounds bounded by upright stones and may belong to a "mixed type" incorporating local traditions from eastern Kazakhstan and the Russian Altai.

Unclassified. In addition to these two main types, we also analyzed one individual from a site with an uncertain burial type: Denj (GUR).

Comparative genomic data are available for several contemporaneous sites in neighboring regions, including: (1) Botai, a horse hunter-herder site dating to ca. 3500 BCE in northern Kazakhstan (Damgaard et al., 2018a); (2) multiple sites of Afanasievo ruminant pastoralists dating to ca. 3000-2500 BCE in the Kazakh and Russian Altai-Sayan region (Allentoft et al., 2015; Narasimhan et al., 2019); (3) Dali, a site in southeastern Kazakhstan whose lowest layers contain a woman dating to ca. 2650 BCE but lacking burial context (Narasimhan et al., 2019); (4) Gonur Tepe, a representative Bactria-Margiana Archaeological Complex (BMAC) site in Turkmenistan dating to 2300-1600 BCE (Narasimhan et al., 2019); and (5) three Lake Baikal sites, Ust'-Ida I, Shamanka II and Kurma Xi, associated with the Glazkovo mortuary tradition and dating to ca. 2200-1800 BCE (Damgaard et al., 2018a; 2018b).

Middle/Late Bronze Age (ca. 1900-900 BCE)

The Middle/Late Bronze Age (MLBA) in Mongolia is characterized by the sudden and widespread appearance of monumental mortuary architecture across Mongolia. Primarily taking the form of stone mounds, but also including stone stelae and other features, these Middle and Late Bronze Age structures remain among the most conspicuous features on the landscape even today. Middle and Late Bronze Age burial mound typology is complex and there is scholarly debate and disagreement on how to precisely define and delineate different mortuary types. In this study, we focused on several main burial forms: Mönkhkhairkhan, Baitag, Deer Stone-Khirigsuur Complex (DSKC), Ulaanzuukh, and Tevsh (Shape). We provide a general overview of these burial types, but acknowledge that not all scholars will agree with all details.

Mönkhkhairkhan (1850-1350 BCE). Dating to after the Chemurchek period, ca. 1850-1350 cal. BCE (Taylor et al., 2019), Mönkhkhairkhan burials are found across northern and western areas of Mongolia and in Tuva, spanning a geographic area approximately 1000 km from west to east and 500 km from north to south. Mönkhkhairkhan burials are characterized by a crouched/flexed burial position, and graves are completely filled in with stones after burial, as seen at the site Ulaan Goviin Uzuur 2 (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). Overlaying the burials are external stone structures consisting of flat round or rectangular platforms. Such barrows are typically 3-5 m in diameter, but occasionally reach up to 40 m in diameter. Ritual structures may include stone circles and stelae. No pottery or animal bones have been reported from within Mönkhkhairkhan burials, and little is known of its economy. Burial goods include tin bronze knives and awls, tin bronze two-trumpet shaped rings, bone spoons, bone arrowheads, and ornaments made of bone, shell, and stone (Clark, 2015; Eregzen, 2016). Knives and rings have analogs both in western Siberia and among the Oijia, Lower Xiajiadian, Siba, and Erlitou cultures of China. Similarities between the grave goods and funerary practices of this culture with those at sites to the west of Lake Baikal (Cis-Baikal) have been previously noted

(Erdenebaatar, 2016; Kovalev, 2017; Kovalev and Erdenebaatar, 2009). We analyzed individuals from two Mönkhkhairkhan sites in this study: Khukh Khoshuunii Boom (KHU) and Ulaan Goviin Uzuur 2 (UAA).

Baitag (1050-900 BCE). Restricted to southwestern Mongolia, Baitag burials consist of non-mounded, small stone rings constructed from a single layer of small flat stone slabs, as seen for example at the site of Uyench (Khovd province, western Mongolia; not analyzed in this study, https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo) (Kovalev and Erdenebaatar, 2009). A central burial pit oriented west-east contains a single individual oriented in a supine position with knees up. Unlike DSKC burials but similar to preceding Altai groups, such as the Mönkhkhairkhan and Chemurchek, the Baitag burials contain various small grave goods, including bronze jewelry. These artifacts share similarities with those included in Karasuk culture graves from the Minusinsk Basin, as well as in burials in Xinjiang and Gansu (Sibu culture) in northwestern China (Kovalev and Erdenebaatar, 2009). In this study, we analyzed one Baitag individual (ULI004) from the site of Uliastai River (ULI), middle terrace.

Deer Stone-Khirigsuur Complex (DSKC) (1350-900 BCE). This culture comprises three different monumental features - *khirigsuurs*, deer stones, and Sagsai-style graves - and is tightly associated with the emergence of horsemanship in the Mongolian Steppe during the late second millennium BCE. In general, DSKC sites are concentrated in the western, northern, and central parts of Mongolia, with only a small number of sites further east (Honeychurch, 2015). *Khirigsuurs* in central Mongolia consist of large stone mounds, surrounded by an outer fence that is either circular or four-cornered in shape, as seen for example at the site of Egiin Gol (Bulgan province, northern Mongolia; not analyzed in this study, https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo) . In the Mongolian Altai, some *khirigsuurs* display stone lines between the central cairn and outer fence, producing a shape that resembles a spoked wheel. Although their exclusive function as burials is a subject of contention (Wright, 2012), *khirigsuurs* often contain a supine human body (Littleton et al., 2012) and do not typically yield other kinds of artifacts. Deer stones are anthropomorphic standing stones found either independently or co-occurring with *khirigsuurs*. Deriving their name from the common motif of stylized deer, carvings on these stelae also depict belts, weapons, and tools - and occasionally even a human face. Many of the weapons depicted on deer stones are of recognizably Karasuk style, bearing a strong resemblance to bronzes found in tombs in the Minusinsk Basin more than 500 km to the northwest (Honeychurch, 2015), and the presence of deer stones in nearby Tuva further support the possibility of long-distance interaction between the Karasuk and the DSKC (Honeychurch, 2015). At many Mongolian *khirigsuurs* and deer stones (particularly in central Mongolia), smaller stone mounds containing the head, jaw, neck, and hooves of individual horses are found surrounding the eastern perimeter of the monument (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). These horse mounds can range in number from a handful into the hundreds or thousands. Osteological study of DSKC horses reveal their use in transport and likely riding, as well as their sophisticated management as herd animals (Taylor et al., 2015; 2018). Another kind of satellite feature found at DSKC sites, open stone circles, often yield partial remains of sheep, goat, or cattle. Analogies in the composition and architecture of deer stone and *khirigsuur* sites with horse sacrifices has led some to interpret deer stones as cenotaphs for people not buried in funerary structures (Kovalev et al., 2014; 2016).

Sagsai-style graves (Taylor et al., 2019; Törbat, 2016) are often associated with the DSKC culture; however, this grave type is not associated with horse sacrifice. Sagsai burials consist of round or rectangular stone platforms without an outer fence, but with large boulders demarcating the edge of the cairn. Beneath the center of the platform, individual burials are positioned within stone cists or narrow pits covered by stone slabs. Individuals are typically arranged in a supine position without burial goods. Alternative names that have been used to describe this burial style include Munguntaiga, Mongun-Taiga, and even *khirigsuur*. A similar burial style known as a "slope burial" due to its common occurrence on the edge of hillslopes is often considered a variant of the Sagsai type. Slope burials consist of a similar stone cairn with four-corner fences and upright corner posts. Sagai-style mounds and slope burials are concentrated in western and northern Mongolia, and representative examples have been excavated in Khövsgöl province (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo).

Radiocarbon modeling dates central Mongolian *khirigsuurs* to between ca. 1350-900 cal BCE, deer stones to ca. 1150-750 cal. BCE, and places the emergence of DSKC horse ritual at ca. 1200 cal BCE (Taylor et al., 2019, 2017). Sagsai-style graves fall within this range (1350-1050 BCE), further strengthening the claim for their affiliation to the DSKC culture sphere (Taylor et al., 2019). It should be noted, however, that these dating estimates could be influenced by taphonomic or dietary processes. In particular, young dates for deer stones may be influenced by radiocarbon contamination (Zazzo et al., 2019), and early dates may be influenced by aquatic reservoir effects. Dairy proteins preserved in dental calculus demonstrate a pastoral, ruminant dairy-based economy at *khirigsuur* and Sagsai sites (Jeong et al., 2018; Wilkin et al., 2020a), and one Sagsai site to date has also yielded evidence of horse milking (Wilkin et al., 2020a). Perhaps buoyed by the innovation or adoption of mounted horseback riding and accompanying changes to the pastoral economy, deer stones and various stone cairns with external fences proliferated over an extremely wide geographic range, reaching modern-day Tuva and southern Russia, Kazakhstan, Kyrgyzstan, and northwest China. We analyzed individuals from four DSKC sites in this study: Arbulag Soum (ARS), Berkh Mountain (BER), Uliastai River Lower Terrace (ULI), and Uushigiin Uver (UUS).

Ulaanzuukh - Tevsh (Shape) (1450-1150 BCE). Beginning in the mid-second millennium BCE, a number of different burial traditions emerged in the southern and southeastern regions of Mongolia. United by a common prone or face-down burial position, these groups were once considered part of the Slab Grave culture, but are now either classified separately as discrete burial types Ulaanzuukh and Tevsh/Shape (Kovalev and Erdenebaatar, 2009) or are sometimes considered a single cultural unit (Ulaanzuukh-Tevsh/Shape) (Honeychurch, 2015). Ulaanzuukh burials (named after the type site), are found within southeast Mongolia and consist of non-mounded square or rectangular platforms surrounded by a wall of upright slabs or layered stone placed over a central burial pit (Dashtseveg et al., 2014). The site of Adgiin Gol (Sukhbaatar province, eastern Mongolia; not analyzed in this study) provides a

representative example of this burial type (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). Tevsh burials, also called Shape burials, are found throughout southern Mongolia and central Inner Mongolia and are similar to Ulaanzuukh burials except that they are hourglass-shaped (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). The walls of Tevsh/Shape burials are typically made of layered stone (masonry), and sometimes with a single edge ringed with upright slabs. Other burial styles in the region, which may represent variant types, include D-shaped and stirrup-shaped burial structures.

Radiocarbon modeling suggests that Ulaanzuukh features date to ca. 1450-1150 BCE, while shape burials could both predate and postdate this mark – although very few have been reliably dated (Taylor et al., 2019). Burials of this culture often contain apparently domestic livestock remains, including sheep, goat, horse, and cattle (Nelson et al., 2009), although the earliest horses from these features date to only ca. 1250 BCE (Taylor et al., 2017). Recent analysis of proteins in human dental calculus has confirmed the utilization of ruminant dairy products and the presence of domestic animals in the Ulaanzuukh economy (Wilkin et al., 2020a). A few bronze knives of Karasuk origin have been found in Ulaanzuukh-Tevsh graves, indicating possible long-distance connections to the Minusinsk basin (Honeychurch, 2015). We analyzed individuals from two Ulaanzuukh sites in this study: Bulgiin Ekh (BUL) and Ulaanzuukh (ULN). We did not analyze individuals from Tevsh/Shape burials in this study.

Unclassified. In addition to these main types, we also analyzed individuals from six Late Bronze Age sites containing burials with uncertain or unclassified cultural affiliations: Biluutiin Am (BIL), Khoit Tsenkher (KHI/KHO), Shar Gobi 3 (SBG), Tsaidam Bag (TSB/TSI), Uliastai River lower terrace (ULI), and Uliastai Zastav II (ULZ). For more information on these burials, see Table S1C.

Comparative genomic data are available for several contemporaneous archaeological sites in neighboring regions, including: (1) four Okunevo sites (Verkhni Askiz, Okunev Ulus, Uybat, Syda 5), dating to 2200-2600 BCE (Allentoft et al., 2015; Damgaard et al., 2018a); (2) five Sintashta sites (Bulanovo, Tanabergen II, Stepnoe VII, and Bol'shekaraganskii, Kamennyi Ambar 5 cemetery), dating to ca. 2200-1700 BCE (Allentoft et al., 2015; Narasimhan et al., 2019); (3) four Central Steppe sites near Krasnoyarsk in western Siberia (Krasnoyarsk Krai, Potroshilovo II, Ust-Bir IV, Chumyash-Perekat-1) dating to 1700-1400 BCE (Narasimhan et al., 2019); (4) three Karasuk sites (Arban I, Sabinka II, and Bystrovka), dating to ca. 1400-1300 BCE (Allentoft et al., 2015).

Early Iron Age (ca. 900-300 BCE)

The Early Iron Age cultures of Inner Asia arose during a time of new technological advancements, including the development of composite bows and the beginnings of iron metallurgy used for items like arrows and horse-riding equipment (Honeychurch, 2015). These cultures include (1) the widespread Slab Grave culture, prevalent in eastern, southeastern, and central Mongolia as well as East Baikal and parts of northern China, and (2) the Sagly/Uyuk and Pazyryk cultures in the Sayan-Altai and portions of northwestern Mongolia. These latter cultures were part of a broader ''Scythian'' cultural phenomenon that spread into eastern Kazakhstan and across the Eurasian steppes, and which was related to Saka groups of northern Iran and the Tian Shan mountains. The Saka were an Iranian group broadly associated with the Scythians. Their later (after 200 BCE) military activities in Sogdia, Bactria, and the Tian Shan were recorded by Persian, Greek, and Chinese sources (Beckwith, 2009). Alongside the technological advancements of the Early Iron Age came increased long-distance interactions and the intensification of grain subsistence outside of the central Mongolian Steppe, but not yet by groups like the Slab Grave culture within Mongolia (Ventresca Miller and Makarewicz, 2019).

Slab Grave (1000-300 BCE). Beginning around 1000 BCE, a new burial style known as Slab Grave began appearing in eastern Mongolia. Slab graves are so called because of the large stone slabs used to mark the surface of the burial and to contain the rectangular burial space (hence in Mongolian they are called ''square burials'') wherein single individuals are interred (Tsybiktarov, 1998). Although occasionally found singly, Slab Grave burials are more typically grouped into small cemeteries (Honeychurch, 2015). Stone slabs are set upright in the ground, and are thus prominent grave markers (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). The burial pits are quite shallow, and human remains are rarely found complete or in good preservation. Over time, the Slab Grave culture expands northward into eastern Baikal and westward into central Mongolia, where it intrudes into former DSKC territory. Some slab graves tear apart the stone structures of *khirigsuurs* to construct the graves, and some even reuse deer stones for standing corner stones or laid-down slabs within the burial pit (Honeychurch, 2015). Unlike earlier Bronze Age burials, grave goods become more common in Slab Grave burials, consisting primarily of bronze beads, buttons, and small ornaments, as well as horse gear, arrowheads, axes, and knives. Stone, ceramic, and bone artifacts are also found in slab graves, and a few burials contained tripod-shaped pottery similar to those from Inner Mongolia and Manchuria or other non-local grave goods such as turquoise and carnelian beads from Central or South Asia (Honeychurch, 2015). Portions of livestock are often set at the edge or just outside of the rectangular burial space. In addition to faunal remains demonstrating the presence of domestic animals in the Slab Grave economy, recent analysis of proteins in human dental calculus has confirmed the utilization of ruminant and horse dairy products (Wilkin et al., 2020a). Although the Slab Grave phenomenon emerges out of the former territory of the Ulaanzuuk culture, archaeological evidence for the relationship between these two groups has been ambiguous. Nevertheless, the similarity of bronze artifacts, especially relating to horse gear and weaponry, found at Slab Grave sites to similar artifacts found in the Altai, Tuva, and Minusinsk regions may indicate a continuation of previously established long-distance relationships between these regions (Honeychurch, 2015). We analyzed individuals from five Slab Grave sites in this study: Bor Bulag (BOR), Morin Tolgoi (MIT), Darsagt (DAR), Shunkhlai Mountain (SHU), and Pesterevo 82 (PTO).

Sagly/Uyuk (500-200 BCE). This Early Iron Age culture centered in the Upper Yenisei River area, in modern-day Tuva, with some extensions into northwestern Mongolia (Murphy, 2003; Savinov, 2002). This culture is also referred to as the Sagly-Bazhy culture, and is best known in Mongolia by the thoroughly excavated site of Chandman Mountain (Ulaangom cemetery) included in this study (Novgorodova et al., 1982; Tseveendorj, 1980) (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). Graves were marked

by a round pile of stones and are often found in cemeteries of one to two dozen graves. Beneath the stone mounds are large log chambers containing several individuals (often assumed to be kin as they include men, women and children) all laid in partially flexed positions on their sides. Portions of sheep are also often placed in the graves. The Sagly/Uyuk log chambers resemble similar log architecture constructed by the contemporaneous Pazyryk culture in the Russian Altai and surrounding areas, and both the Sagly/Uyuk and Pazryk have been associated with the broader Saka culture (Parzinger, 2006). Similar to Slab Graves, recent analysis of proteins in human dental calculus has confirmed the utilization of ruminant and horse milk among those at Chandman Mountain (Wilkin et al., 2020a). Isotopic studies have also shown that some Uyuk communities, including at Chandman Mountain, had a significant amount of millet in their diet (Murphy et al., 2013; Wilkin et al., 2020b). This links them to agropastoralist cultures of the Western Steppe, where the intensification of millet cultivation occurred during the second millennium BCE (Ventresca Miller and Makarewicz, 2019). We analyzed individuals from one Sagly/Uyuk site in this study: Chandman Mountain (CHN).

Pazyryk (500-200 BCE). This culture is known mainly for its type site of Pazyryk, whose large tombs contain numerous exotic imports, including silks from China and textiles from Achaemenid Persia (Rudenko, 1970). Pazryk burials are found mostly within the northern Altai areas of Russia, far eastern Kazakhstan (Samashev, 2011) and northwestern Mongolia (Törbat et al., 2009). Similar to Sagly/Uyuk and other 'Saka' style graves, Pazyryk burials are marked by round piles of stones. Beneath these stone piles, however, most Pazyryk graves have smaller wooden chambers with only one or two persons; their size and burial goods vary greatly, though many of them are accompanied by whole horses laid beside the burial chamber (Kubarev and Shul'ga, 2007) (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). No new Pazyryk individuals were included in this study; however, they are important to consider because the northern Altai practice of whole horse burials later appears in scattered central Mongolia cemeteries of the subsequent Xiongnu period. Genome-wide data from Pazyryk individuals have been previously reported from site of Berel in Altai region of Kazakhstan (Unterländer et al., 2017).

Comparative genomic data are available for several contemporaneous sites in neighboring regions, including: (1) the early Sarmatian site Pokrovka in southwestern Russia, dating to ca. 500-100 BCE (Unterländer et al., 2017), a Scythian individual from the Samara region dating to ca. 300 BCE (Mathieson et al., 2015), and nine Sarmatian sites in southwestern Russia (Chebotarev V, Kamyshevahsky X, Nesvetay II, Nesvetay IV and Tengyz), northern Kazakhstan (Bestamak and Naurzum Necropolis), and the southern Ural region (Cherniy Yar and Temyaysovo) (Damgaard et al., 2018b; Krzewińska et al., 2018); (2) the Pazyryk site of Berel in the Altai, dating to ca. 400-200 BCE (Unterländer et al., 2017); (3) the Saka sites of Borli, Karasjok-1, Karasjok-6, Nazar-2, Sjartas (Zjartas), and Taldy-2 in Kazakhstan (Damgaard et al., 2018b), and the sites of Basquiat I, Keden, and Ornek in the Tian Shan (Damgaard et al., 2018b); and (4) the Tagar site of Grishkin Log 1 in the Minusinsk Basin (Damgaard et al., 2018b). Data from three other potentially relevant sites (the Aldy-Bel site Arzhan 2 in Tuva, dating to ca. 700-500 BCE, and the Zevakino-Chilikta sites Ismailovo and Zevakino in eastern Kazakhstan, dating to ca. 900-600 BCE; (Unterländer et al., 2017) were excluded from analysis due to insufficient genetic coverage for comparison.

Xiongnu (ca. 200 BCE to 100 CE)

During the late first millennium BCE, a radically new multi-regional political entity formed in Mongolia, known as the Xiongnu empire. The Xiongnu empire is attested not only by historical records but also by ample archaeological remains throughout Inner Asia (Brosseder and Miller, 2011; Honeychurch, 2015). For roughly three centuries the Xiongnu ruled from their core realms in central and eastern Mongolia, expanding into western Mongolia, northern China and eastern Baikal, as well as making inroads into more distant regions in Central Asia. Most graves of the Xiongnu period were shaft pits set beneath thick rings of stones on the surface. These burials represent the vast network of regional and local elites and not the ''commoner'' people of Xiongnu society, whose burials are far less conspicuous, lying under small piles of stones or in unmarked pits. The graves of the uppermost ruling elites of the empire, on the other hand, were constructed on a far grander scale than that of ring graves.

While ring grave structures are found throughout the entire Xiongnu era, prestige accoutrements (and to some degree burial rituals) changed during the course of the empire. According to these changes, we can discern a general division between Early (200-50 BCE) and Late (50 BCE - 100 CE) Xiongnu periods (Miller, 2014). Overall, Xiongnu graves are marked by a dramatic increase in grave goods and furnishings as compared to previous time periods and cultures in Mongolia. As the Xiongnu expanded their empire, they conquered numerous neighboring groups to their east and west as well as subduing their Han Dynasty neighbors to the south (Di Cosmo, 2002). They continually traded and warred with Han China, defying the Great Wall boundaries, and held significant sway over the Silk Road kingdoms of Central Asia (Hulsewé, 1979). The findings of exotic items from China, Persia and the Mediterranean attest to these far-flung interactions, with Egyptian-style faience beads in graves of local elites and Roman glass bowls in the tombs of the rulers (Miller and Brosseder, 2017; Eregzen, 2011). The end of the Xiongnu period ca. 100 CE is marked by the widespread decline of Xiongnu power and influence following defeats by the Xianbei in northeastern China and the Han Dynasty of China, although isolated groups from the Xiongnu empire continued to exist in northern China until the 5th century CE.

Early Xiongnu (200-50 BCE). Prestige items during the Early Xiongnu period are dominated by large bronze belt pieces; however, burial customs within graves of the Early period varied to a great degree between regions. One example of this occurs at Salkhityn Am cemetery, where rituals of ring graves show a high degree of variation, even including offerings of whole horses that are more typical of Altai elites such as those in Pazyryk graves (Ölziibayar et al., 2019) (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). We analyzed individuals from three Early Xiongnu sites in this study: Astyn Gol (AST), Buural Uul (BAU/BRL/BUU), and Salkhityn Am (SKT).

Late Xiongnu (50 BCE - 100 CE). Prestige items in the Late Xiongnu period shift to more iron items, often covered with gold foil or even inlaid with precious stones, and increasingly focused on long-distance exotic materials. At the same time, burial customs in ring graves throughout the empire become more regularized. Most elites were buried in wooden coffins in shaft pits with livestock portions and ceramic vessels set beside the coffin (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). During the Late period, the high ruling Xiongnu elites adopted a radically new form of burial structure. These square tombs were marked on the surface by rectangular stone structures with trapezoidal 'ramp' entryways, their burial pits were extremely deep, and wooden coffins were decorated and nested within larger wooden chambers (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). We analyzed individuals from 26 Late Xiongnu sites in this study: Atsyn Am (ATS), Baruun Mukhdagiin Am (BAM), Baruun Khovdiin Am (BRU), Burkhan Tolgoi (BTO), Chandman Mountain (CHN), Delgerkhaan Uul (DEL), Khanan Uul (DOL), Duulga Uul (DUU); Emeel Tolgoi (EME), Khudgiin Am (HUD), Ikh Tokhoirol (IKT), Il'movaya Pad (IMA), Jargalantyn Am (also called Jargalantyn Khondii; JAA/JAG), Tarvagatain Am (also called Khoit Tsenkher; KHO), Naimaa Tolgoi (NAI), Sant Uul (SAN), Solbi Uul (SOL), Songino Khairkhan (SON), Takhityn Khotgor (TAK), Tavan Tolgoi (TAV), Tevsh Mountain (TEV), Ulaanzuukh (ULN), Ovgont (UVG), Yuroo II (YUR), Tamiryn Ulaan Khoshuu (also called Burkhan Tolgoi; BUR/TMI/TUH/TUK), and Uguumur Uul (UGU).

Comparative genomic data are available for a few contemporaneous sites in neighboring regions, including: (1) two early Xiongnu individuals from Khövsgöl (Hovsgol) province dating to 50-350 BCE (Damgaard et al., 2018b); (2) a late Xiongnu royal tomb (DA39.SG) in Arkhangai dating to 80-160 CE (Damgaard et al., 2018b).

Early Medieval (ca. 100-850 CE)

After the fall of the Xiongnu, Xianbei groups from northeast China pushed into Mongolia, although historical and archaeological evidence for the establishment of large and long-lasting Xianbei polities appears only in northern China, not in Mongolia (Miller, 2016). One individual in this study (TUK001) at the site of Tamiryn Ulaan Khoshuu (Burkhan Tolgoi) dates to the era of Xianbei power in Inner Asia; however, there is no cultural context that could affirm affiliation with the Xianbei or other groups of northeastern China. Instead, recent excavations at this site have yielded artifacts, such as pottery from the Kwarezm oasis cultures near the Aral Sea and coins of the Sassanian Persian empire, that indicate significant interactions with areas in Central Asia and much farther west. In the mid-fourth century, a large polity known as the Rouran purportedly took over all of Mongolia; however, there is little recorded history about the Rouran (Kradin, 2005), and only one grave found so far can be dated to the Rouran era (Li et al., 2018; Nei Menggu zizhiqu wenwu kaogu yanjiusuo and International Institute for the Study of Nomadic Civilizations, 2015). The archaeology of the second to sixth centuries in Mongolia, i.e., the Xianbei and Rouran eras, constitute an extremely new field of research (Odbaatar and Egiimaa, 2018).

The most prominent political entities in the Early Medieval era are the Türk and Uyghur empires, the latter being an immediate dynastic takeover from the former. Numerous burials of the Türk era have been unearthed in Mongolia. By contrast, far fewer Uyghur burials have been identified and excavated to date.

Türk (550-750 CE). Göktürkic tribes of the Altai Mountains established a political structure across Eurasia beginning in 552 CE, with an empire that ruled over Mongolia from 581-742 CE (Golden, 1992). A brief period of disunion occurred between 659-682 CE, during which the Chinese Tang dynasty laid claim over Mongolia. One individual from this study (TUM001) was a sacrificial person within the ramp of a Chinese-style tomb in central Mongolia dating (via tomb inscription) to this exact time period. The other Türkic era individuals in this study were excavated from conventional Türkic style graves. Features of the Türk period include numerous stone statues and stone offering boxes across the steppe landscape, while burials are often arranged as small groups of graves or single graves inserted into burial grounds of earlier Bronze to Iron ages. Most elites were interred within wooden coffins as single individuals buried beneath a stone mound, and many were buried with whole horses equipped with riding gear (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). Other burials were in small wooden coffins without whole horses beside them. We analyzed individuals from 5 Türk sites in this study: Nomgonii Khundii (NOM), Shoroon Bumbagar (Türkic mausoleum; TUM), Zaan-Khoshuu (ZAA), Uliastai River Lower Terrace (ULI), and Umuumur uul (UGU).

Uyghur (750-850 CE). In the mid-eighth century, Uyghur tribes from the Upper Yenesei region overthrew the Türk rulers and immediately established a Mongolia-based empire, taking over the Orkhon valley as their capital and establishing a dynasty from 744-840 CE (Mackerras, 1972). Most Uyghur period burials excavated to date, including those from the Olon Dov burial ground (OLN) included in this study, lie in the vicinity of the Kharbalgas capital in the Orkhon Valley. Most of the burials excavated were discovered beneath large earthen enclosures that contained ritual structures for venerating the uppermost elites. These conspicuous ritual enclosures occur as single monuments or in small groups, and they are found in several locations throughout the foothills of the nearby the Uyghur capital. These monumental tombs with ramp entries and vaulted brick chambers were likely reserved for the ruling nobility of the Uyghur empire (Odbaatar, 2016) (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). One individual in this study (OLN006) was found in a monumental tomb. A second, more modest category of Uyghur burials consists of stone structures placed on the surface, either square or round in shape, that contain multiple individuals (Erdenebat 2016) (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). Dozens of these burials have been documented at Olondov (Erdenebat et al., 2012), and most of the Uyghur individuals in this study are from such graves. One such grave at Olon Dov, grave 19, contained the remains of multiple individuals, six of whom are included in this study. Other scattered examples of single Uyghur graves have been found in Mongolia, and we analyzed one of these (ZAA001) from the site of Zaan-Khoshuu. Although a few large 'royal' complexes have been found elsewhere in central Mongolia, no significant cemeteries outside the capital region have yet been found. We analyzed individuals from two Uyghur sites in this study: Olondov (OLN) and Zaan-Khoshuu (ZAA).

Comparative genomic data are available for contemporaneous sites in neighboring regions, including: (1) Alan sites in North Ossetia-Alania and Alan 51 from the Caucasus (Damgaard et al., 2018b); (2) the Rouran site of Khermen Tal site from Arkhanggai, Mongolia (Li et al., 2018).

Late Medieval (ca. 850-1650 CE)

This period in Mongolia is dominated mostly by the power struggles of two empires established by the Khitans (907-1125 CE) and the Mongols (1206-1368 CE). Burials from the Khitan era are virtually unknown in Mongolia, whereas numerous graves from the Mongol era have been documented and unearthed. So-called cave burials are known from both periods (Bemmann and Nomguunsüren, 2012), but their human remains were not included in this study.

Khitan (ca. 900-1100 CE). After the collapse of the Uyghur empire in Mongolia in 840 CE, the Khitans of northeast China established the powerful Liao Dynasty in 916 CE. Although based in Manchuria, the Khitans conquered and controlled the steppe of present-day Mongolia through a system of garrisons and long walls, deporting people from other conquered regions, such as northern Korea, to Mongolia (Kradin and Ivliev, 2008). The dissolution of the Khitan empire in 1125 CE led to a power vacuum in Mongolia until the rise of Chinggis Khan in the early 13th century CE. To date, very few Khitan era graves have been found in Mongolia. The site of Ulaan Kherem II (ULA) has yielded one Khitan-era grave (ULA001), and two Khitan-era unmarked graves of a man and woman were also discovered during the excavation of a Xiongnu settlement at Zaan Khoshuu (ZAA) beneath an older collapsed building (Nei Menggu zizhiqu wenwu kaogu yanjiusuo and International Institute for the Study of Nomadic Civilizations, 2015; Ochir et al., 2016). The man, found in a pit within the pit-house, was buried in a simple pit with a quiver and arrows. The woman, found nearby a pit-house, was buried in full dress and placed in a supine position with her head to the northwest inside a wooden coffin, along with pottery of the Khitan era (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). These burials are significantly different in form and structure from other Khitan burials in northern China, where the core of the empire was located. At present, no monumental tombs of high Khitan elites have been found in Mongolia.

Mongol (ca. 1200-1400 CE). The home base of the Mongol tribe was in the forest-steppe zone at the Onon and Kerülen (Kherlen) rivers in northeastern Mongolia. From this core region they successfully conquered the Eurasian steppes and most of their sedentary neighbors in the adjacent regions. Historical records indicate that they transferred a large number of defeated people, war captives and slaves all over their growing empire; they also fostered trade, the exchange of knowledge, techniques, and technicians (Allsen, 2015). Mongol burials are typically situated in small groups on flat southern slopes or placed within Bronze and Iron Age cemeteries. They are marked above ground with stones in an irregular, flat, oval or rounded, one-layered setting (https://edmond.mpdl.mpg.de/imeji/collection/2ZJSw35ZTTa18jEo). The pit is normally between 50-150 cm deep, rarely deeper, and very seldom constructed as a niche. Typical Mongol burials contain one person placed in a supine position and sometimes in a wooden coffin, with the head to the north. A very characteristic feature of Mongol burials is the inclusion of a tibia from small livestock, mostly sheep, placed near the head and sometimes in a vessel. There are two ideal burial types concerning grave goods: one equipped with bow, arrow, quiver, horse equipment, and belt with attachments, and a second with scissors, a comb, a mirror, beads and a *bogtag* – a long hat made out of birch bark, covered with silk and decorated with golden ornament. Graves of these standard types are spread all over Mongolia, and at present no regional differences have been reported and no monumental burials are known (Erdenebat, 2009; Lkhagvasüren, 2007). The Mongol burials included in this study are of these types, which consist of the burials of local steppe warriors and elites of the Mongol empire. Individuals from the cosmopolitan capital of Karakorum were not sampled in this study.

Historical records mention a large amount of foreign people who migrated, whether for opportunity or by force, into the core steppe regions of the Mongol empire (Allsen, 2015). Given the intriguing results of extreme genetic diversity among local elite constituents for the Xiongnu era, one might expect a similar or even greater diversity during the Mongol era. However, within the core steppe realms, lower local levels of the Mongol empire appear not to have been as open. The supposed mass of incoming foreigners must be sought in other burial contexts, not those of Mongol tradition.

Unclassified. In addition to these main types, we also analyzed individuals from three Late Medieval sites containing burials with uncertain or unclassified cultural affiliations: Shunkhlai Mountain (SHU), Tsaidam Bag (TSB/TSI), Uushigiin Uver (UUS).

Because no comparative genomic data are available for contemporaneous sites, we compared our Late Medieval data to modern Mongolic speaking populations (Buryat, Khamnegan, Kalmyk, Mongol, Daur, Tu, Mongola) (Jeong et al., 2019; Lazaridis et al., 2014; Patterson et al., 2012).

METHOD DETAILS

Radiocarbon dating of sample materials

A total of 30 new radiocarbon dates were obtained by accelerator mass spectrometry (AMS) of bone and tooth material at the Curt-Engelhorn-Zentrum Archäometrie (CEZA) in Mannheim, Germany (n = 28) and the University of Cologne Centre for Accelerator Mass Spectrometry (CologneAMS) (n = 2). Selection for radiocarbon dating was made for all burials with ambiguous or unusual burial context and for all individuals appearing as genetic outliers for their assigned period. Uncalibrated direct carbon dates were successfully obtained for all bone and tooth samples (Table S4). An additional 74 previously published radiocarbon dates for individuals in this study were also compiled and analyzed, making the total number of directly dated individuals in this study to 98 (104 total dates). Dates were calibrated using OxCal v.4.3.2 (Ramsey, 2017) with the r:5 IntCal13 atmospheric curve (Reimer et al., 2013).

Of the 104 total radiocarbon dates analyzed in this study, 25 conflicted with archaeological period designations reported in excavation field notes or previous publications (Table S4). Four burials of uncertain cultural context were successfully assigned to the Middle/Late Bronze Age (BIL001, MIT001) and Late Medieval periods (UUS002, ZAA003). One burial originally assigned to the Late Medieval period was reassigned to the pre-Bronze Age following radiocarbon dating (ERM001), and one burial originally assigned to the Middle/Late Bronze Age was similarly reassigned to the Early Bronze Age (IAG001). This suggests that early burials may be underreported in the literature because they are mistaken for later graves. Likewise three burials originally classified as Late Medieval were found to be hundreds or thousands of years older, dating to the Early Medieval (TSB001) and Middle/Late Bronze Age (ULZ001, TSI001) periods. Although some highly differentiated burial forms can be characteristic of specific locations and time periods, simple burial mounds also exist for all periods and - lacking distinctive features - they can be difficult if not impossible to date without radiometric assistance.

In addition to early burials being mistaken for later ones, late burials were also misassigned to earlier periods. For example, three burials originally assigned to the Middle/Late Bronze Age were determined to date to the Early Iron Age (DAR001), Xiongnu (ULN004), and Late Medieval (SHU001) periods, and two Early Iron Age (CHN010, CHN014), six Xiongnu (TUK001, UGU001, DUU002, BRL001, BAU001, DEE001), and two Early Medieval (ULA001, ZAA005) graves were likewise reassigned to later periods following radiocarbon dating. Part of the difficulty in correctly assigning archaeological period to later burials relates to the frequent reuse of earlier graves and cemeteries by populations from later periods. The site reports of several Xiongnu excavations noted burial intrusions, displaced burials, and other indications of burial disturbance and reuse. However, evidence of burial reuse may also be subtle and easily overlooked. As such, we recommend great care in making cultural or temporal assignments at multi-period cemeteries or for any burials showing evidence of disturbance.

Sampling for ancient DNA recovery and sequencing

Sampling was performed on a total of 169 teeth and 75 petrosal bones from fragmented crania originating from 225 individuals (Table S1C). For 14 individuals, both a tooth and a petrosal bone were sampled (Table S2B). For three individuals, two teeth were sampled, and for one individual, two teeth and one petrosal bone were sampled (Table S2B). For Mongolian material, whole teeth and petrosal bone (except ERM) were collected at the physical anthropology collections of the National University of Mongolia and the Institute for Archaeology and Ethnology under the guidance and supervision of M. Erdene and S. Ulziibayar. Petrosal and tooth material from ERM were provided by J. Bemmann. For Russian material, whole teeth alongside petrosal bone or bone were collected from the Institute for Mongolian, Buddhist, and Tibetan Research as well as the Buryat Scientific Center, Russian Academy of Sciences (RAS). After collection, the selected human skeletal material was transferred to the Max Planck Institute for the Science of Human History (MPI-SHH) for genetic analysis.

Laboratory procedures for genetic data generation

Genomic DNA extraction and Illumina double-stranded DNA (dsDNA) sequencing library preparation were performed for all samples in a dedicated ancient DNA clean room facility at the MPI-SHH, following published protocols (Dabney et al., 2013) with slight modifications (Mann et al., 2018). We applied a partial treatment of the Uracil-DNA-glycosylase (UDG) enzyme to confine DNA damage to the ends of ancient DNA molecules (Rohland et al., 2015). Such "UDG-half" libraries allow us to minimize errors in the aligned genetic sequence data while also maintaining terminal DNA misincorporation patterns needed for DNA damage-based authentication. Library preparation included double indexing by adding unique 8-mer index sequences at both P5 and P7 Illumina adapters. After shallow shotgun sequencing for screening, we enriched libraries of 195 individuals with $\geq$ 0.1% reads mapped on the human reference genome (hs37d5; GRCh37 with decoy sequences) for approximately 1.24 million informative nuclear SNPs ("1240K") by performing an in-solution capture using oligonucleotide probes matching for the target sites (Mathieson et al., 2015). In addition, eight samples (see Tables S2A and S2B) were also selected and built into single-stranded DNA (ssDNA) sequencing libraries for comparison. Single-end 75 base pair (bp) or paired-end 50 bp sequences were generated for all shotgun and captured libraries on the Illumina HiSeq 4000 platform following manufacturer protocols. Output reads were demultiplexed by allowing one mismatch in each of the two 8-mer indices.

QUANTIFICATION AND STATISTICAL ANALYSIS

Sequence data processing

Short read sequencing data were processed by an automated workflow using the EAGER v1.92.55 program (Peltzer et al., 2016). Specifically, in EAGER, Illumina adaptor sequences were trimmed from sequencing data and overlapping sequence pairs were merged using AdapterRemoval v2.2.0 (Schubert et al., 2016). Adaptor-trimmed and merged reads with 30 or more bases were then aligned to the human reference genome with decoy sequences (hs37d5) using BWA aln/samse v0.7.12 (Li and Durbin, 2009). A non-default parameter "-n 0.01" was applied. PCR duplicates were removed using dedup v0.12.2 (Peltzer et al., 2016). Based on the patterns of DNA misincorporation, we masked the first and last two bases of each read for UDG-half libraries and 10 bases for non-UDG single-stranded libraries, using the trimbam function in bamUtils v1.0.13 (Jun et al., 2015), to remove deamination-based 5′ C>T and 3′ G>A misincorporations. Then, we generated pileup data using samtools mpileup module (Li and Durbin, 2009), using bases with Phred-scale quality score $\geq$ 30 ("-Q30") on reads with Phred-scale mapping quality score $\geq$ 30 ("-q30")

from the original and the end-masked BAM files. Finally, we randomly chose one base from pileup for SNPs in the 1240K capture panel for downstream population genetic analysis using the pileupCaller program v1.2.2 (https://github.com/stschiff/sequenceTools). For C/T and G/A SNPs, we used end-masked BAM files, and for the others we used the original unmasked BAM files. For the eight ssDNA libraries, we used end-masked BAM files for C/T SNPs, and the original BAM files for the others.

In cases where more than one sample was genetically analyzed per individual, we compared the amount of human DNA between samples. For pairs of petrous bone and teeth, human DNA was higher in the petrous bone in 8 of 13 individuals, and higher in the teeth of 5 of 13 individuals (Table S2B). In addition, intra-individual sample variation was high, as evidenced by the high variance observed between paired tooth samples (Table S2B). Finally, in a comparison of dsDNA and ssDNA libraries, ssDNA libraries yielded a higher endogenous content in 7 of 8 library pairs. All data from paired samples were merged prior to further analysis.

Of the 225 new individuals analyzed, 18 failed to yield sufficient human DNA (< 0.1%) on shotgun screening (Table S2A) and a further 6 individuals failed to yield at least 10,000 SNPs after DNA capture (Table S2A). These 24 individuals were excluded from downstream population genetic analysis.

Data quality authentication

To confirm that our sequence data consist of endogenous genomic DNA from ancient individuals with minimal contamination, we collected multiple data quality statistics. First, we tabulated 5′ C>T and 3′ G>A misincorporation rate (Table S2A) as a function of position on the read using mapDamage v2.0.6 (Jónsson et al., 2013). Such misincorporation patterns, enriched at the ends due to cytosine deamination in degraded DNA, are considered as a signature of the presence of ancient DNA in large quantities (Sawyer et al., 2012). Second, we estimated mitochondrial DNA contamination for all individuals using the Schmutzi program (Renaud et al., 2015). Specifically, we mapped adaptor-removed reads to the revised Cambridge Reference Sequence of the human mitochondrial genome (rCRS; NC_012920.1), with an extension of 500 bp at the end to preserve reads passing through the origin. We then wrapped the alignment to the circular reference genome using circularmapper v1.1 (Peltzer et al., 2016). The contDeam and schmutzi modules of the Schmutzi program were successively run with the world-wide allele frequency database from 197 individuals, resulting in estimated mitochondrial DNA contamination rates for each individual (Table S2A). Last, for males, we also estimated the nuclear contamination rate (Table S2A) based on X chromosome data using the contamination module in ANGSD v0.910 (Korneliussen et al., 2014). For this analysis, an increased mismatch rate in known SNPs compared to that in the flanking bases is interpreted as the evidence of contamination because males only have a single copy of the X chromosome and thus their X chromosome sequence should not contain polymorphisms. We report the Method of Moments estimates using the "method 1 and new likelihood estimate," but all the other estimates provide qualitatively similar results.

Ten individuals were estimated to have > 5% DNA contamination (mitochondrial or X) or uncertain genetic sex (Table S2A); these individuals were excluded from downstream population genetic analysis.

Genetic sex typing

We calculated the genetic sequence coverage on the autosomes and on each sex chromosome in order to obtain the ratio between the sex chromosome coverage and the autosome coverage. For 1240K capture data, we observe females to have an approximately even ratio of X to autosomal coverage (X-ratio of ~0.8) and a Y-ratio of 0, and males to have approximately half the coverage on X and Y as autosomes (~0.4). Genetic sex could be determined for a total of 224 individuals, of which 100 were female and 124 were male (Table S2C).

Uniparental haplogroup assignment

We called mitochondrial consensus sequence from the Schmutzi output using the log2fasta program in the Schmutzi package, with quality threshold of 10. We then assigned each consensus sequence into a haplogroup (Table S2C) using the HaploGrep 2 v2.1.19 (Weissensteiner et al., 2016). For the Y haplogroup assignment, we took 13,508 Y chromosome SNPs listed in the ISOGG database and made a majority haploid genotype call for each male using pileupCaller (with "-m MajorityCalling" option). We assigned each individual into a haplogroup (Table S2C) using a patched version of the yHaplo program (Poznik, 2016) downloaded from https://github.com/alexhbnr/yhaplo. This version takes into account high missing rate of aDNA data to prevent the program from stopping its root-to-tip haplogroup search prematurely at an internal branch due to missing SNP and therefore assigning a wrong haplogroup. We used "–ancStopThresh 10" following the developer's recommendation. Haplogroup assignments are shown in Figures S2A and S2B.

Estimation of genetic relatedness

To evaluate the relatedness within our dataset, we calculated pairwise mismatch rate of haploid genotypes on automosomes across all individuals. The pairwise mismatch rate for each pair of individuals, is defined as the number of sites where two individuals have different alleles sampled divided by the total number of sites that both individuals have data. The pairwise mismatch rate between unrelated individuals is set as the baseline and the coefficient of relationship is inversely linear to the baseline pairwise mismatch rate. More detailed description can be found in the Supplemental Materials of (Jeong et al., 2018).

A total of 15 first or second degree genetic relationships were observed across the dataset (Table S2D), of which 10 date to the Xiongnu era. Additionally, in one case, a tooth and petrosal bone thought to belong to one individual (AT-871) were later discovered

to belong to two different individuals (OLN001.A and OLN001.B). In another case, two teeth (AT-728 and AT-729) thought to belong to different individuals were found to originate from the same individual (TUK001/TAV008).

Data filtering and compilation for population genetic analysis

To analyze our dataset in the context of known ancient and modern genetic diversity, we merged it with previous published modern genomic data from i) 225 worldwide populations genotyped on the Human Origins array (Jeong et al., 2019; Lazaridis et al., 2014), ii) 300 high-coverage genomes in the Simons Genome Diversity Project ("SGDP") (Mallick et al., 2016), and iii) currently available ancient genomic data across Eurasian continent (Allentoft et al., 2015; Damgaard et al., 2018a; 2018b; Fu et al., 2014; 2016; Haak et al., 2015; Haber et al., 2017; Harney et al., 2018; Jeong et al., 2016; 2018; Jones et al., 2015; Kılınç et al., 2016; Lazaridis et al., 2016, 2017; Mathieson et al., 2015; 2018; McColl et al., 2018; Narasimhan et al., 2019; Raghavan et al., 2014; 2015; Rasmussen et al., 2010; 2014; 2015; Sikora et al., 2019; Unterländer et al., 2017; Yang et al., 2017). We obtained 1,233,013 SNP sites (1,150,639 of which on autosomes) across our dataset when intersecting with the SGDP dataset, and 597,573 sites (593,124 of which on autosomes) when intersecting with the Human Origins array.

Analysis of population structure and relationships

We performed principal component analysis (PCA) on the merged dataset with the Human Origins data using the smartpca v16000 in the Eigensoft v7.2.1 package (Patterson et al., 2006). Modern individuals were used for calculating PCs (Figure S3A), and ancient individuals were projected onto the pre-calculated components using "lsqproject: YES" option (Figure 2; Figure S3B). To characterize population structure further, we also calculated *f3* and *f4* statistics using qp3Pop v435 and qpDstat v755 in the admixtools v5.1 package (Patterson et al., 2012). We added "*f4mode: YES*" option to the parameter file for calculating *f4* statistics.

Admixture modeling using qpAdm

For modeling admixture and estimating ancestry proportions, we applied qpWave v410 and qpAdm v810 in the the admixtools v5.1 package (Patterson et al., 2012) on the merged dataset with the SGDP data to maximize resolution. To model the target as a mixture of the other source populations, qpAdm utilizes the linearity of *f4* statistics, i.e., one can find a linear combination of the sources that is symmetrically related to the target in terms of their relationship to all outgroups in the analysis. qpAdm optimizes the admixture coefficients to match the observed f4 statistics matrix, and reports a *p-value* for the null hypothesis that the target derives their ancestry from the chosen sources that are differently related to the outgroups (i.e., when $p < 0.05$, the null hypothesis is rejected so that the target is different from the admixture of chosen sources given the current set of outgroups). The chosen outgroups in qpAdm needs to be differentially related to the sources such that a certain major ancestry is "anchored" in the test, which is rather heuristic. We used qpWave to test the resolution of a set of outgroups for distinguishing major ancestries among Eurasians, as well as the genetic cladility between populations given a set of outgroups. We used a set of eight outgroup populations in our study: Central African hunter-gatherers Mbuti.DG (n = 5), indigenous Andamanese islanders Onge.DG (n = 2), Taiwanese Aborigines Ami.DG (n = 2), Native Americans Mixe.DG (n = 3), early Holocene Levantine hunter-gatherers Natufian (n = 6) (Lazaridis et al., 2016), early Neolithic Iranians Iran_N (n = 8) (Lazaridis et al., 2016; Narasimhan et al., 2019), early Neolithic farmers from western Anatolia Anatolia_N (n = 23) (Mathieson et al., 2015), and a Pleistocene European hunter-gatherer from northern Italy Villabruna (n = 1) (Fu et al., 2016).

To evaluate potential sex bias (Figure S2C), we applied qpAdm to both the autosomes (default setting) and the X chromosome (adding "chrom:23" to the parameter file) for comparing the difference in the estimated ancestry proportions. For a certain ancestry, we calculated sex-bias Z score using the proportion difference between P_A and P_X divided by their standard errors ($Z = (P_A - P_X / \sqrt{\sigma_A^2 + \sigma_X^2})$, where σ_A and σ_X are the corresponding jackknife standard errors, as previously performed in (Mathieson et al., 2018). Therefore a positive Z score suggests autosomes harbor a certain ancestry more than X chromosomes do, indicating male-driven admixture. A negative Z score, in contrast, suggests female-driven admixture. The qpAdm estimates from both autosomes and the X chromosome are available in Table S5K.

Dating admixture events via DATES

We used DATES v753 (Narasimhan et al., 2019) to estimate the time of admixture events in ancient individuals (Figure S6B), and convert the estimated admixture date in generation into years assuming 29 years per generation (Patterson et al., 2012). We show the admixture dates in years before present (Figure S6A) by adding the age of each ancient population (i.e., mean value of the midpoint of the 95% confidence interval of available calibrated 14C dates in each population). The standard error of DATES estimates come from the weighted block jackknife, an option in DATES parameter file. In the parameter file for running DATES, we used "binsize: 0.001," "maxdis: 1," "runmode: 1," "mincount: 1," "lovalfit: 0.45" in every run, same to the example file at https://github.com/priyamoorjani/DATES/blob/master/example/par.dates.

Phenotypic SNP analyses

We examined 49 SNPs in 17 genes (Table S2E) known to be associated with phenotypic traits or with positive selection in Eurasia (Jeong et al., 2018). Given the low coverage of ancient DNA data, we focused on five of these genes and calculated the likelihood of allele frequency for SNPs in each ancient population based on the counts of reads covering on the SNP following a published strategy

(Mathieson et al., 2015). In the allele frequency calculation, we classified all ancient individuals before Middle/Late Bronze Age into a single group, and kept three genetic groups during MLBA (Khövsgöl_LBA, Altai_MLBA, Ulaanzuukh), two genetic groups during Iron Age (Chandman_IA, SlabGrave), one group for Xiongnu, one group for Early Medieval and one group for Late Medieval. We calculated allele frequency at five loci (Table S2E) that are associated with lactase persistence (LCT/*MCM6*), skin pigmentation (*OCA2*, *SLC24A5*), alcohol metabolism (*ADH1B*), and epithelial phenotypes including shovel-shaped incisor (*EDAR*) (Figure 5).

Genetic clustering of ancient individuals into analysis units

To further characterize the dynamic changes of the Eastern Steppe gene pools using group-based analyses, we quantitatively examined genetic differences among the analyzed individuals in combination with their temporal, archeological, and geographic information. We first obtained an approximate map of population structure by observing the position of ancient individuals on the PCA calculated from 2,077 present-day Eurasian individuals. PC1 separates geographically eastern and the western populations, PC2 captures the internal variations in eastern Eurasians, and PC3 captures variations in western Eurasians, thus allowing us to characterize an overall pattern of genetic changes through time and helping us to formulate explicit hypotheses regarding the genetic relationships between groups and individuals. Second, we computed outgroup-*f3* and symmetric-*f4* statistics to (1) quantify genetic similarity between individuals/groups falling together on PCA and (2) explore populations whose ancestry through admixture may have contributed to the differences observed between pairs of groups. Third, we identified representative ancient populations to serve as proxies for five distinct ancestries that we then further investigate (Table S3B). We changed the specific ancestry proxy for our test groups based on the temporal and archeological records accordingly. Using these ancestry proxies, we performed a formal admixture modeling using qpWave/qpAdm, which tests the difference between the target and a combination of the proxies (i.e., an admixture model) with regard to their genetic affinity to outgroups. We applied the same admixture models for test groups belonging to the same time/culture/geography category to compare them in a straightforward manner (Figures 3 and 4). In the following paragraphs, we describe each of the genetics-based analysis groups reported in our dataset, as well as the principles we applied to model their genetic ancestry using qpAdm.

Pre-Bronze Age

- New genetic groups: eastMongolia_preBA(1), centralMongolia_preBA(1), and Fofonovo_EN(4)
- Published genetic groups: DevilsCave_N(6), and Baikal_EN(9)

Our dataset adds three Ancient Northeast Asian (ANA)-related genetic groups before the start of the Bronze Age in eastern Eurasia. During this period, we observe the wide distribution of this ANA ancestry from Lake Baikal to the Russian Far East, spanning more than 2,000 km. As Baikal_EN has been modeled to have ~10% Ancient North Eurasian (ANE) ancestry, we also investigated the possible genetic contribution from ANE in our pre-Bronze Age Mongolian and Baikal groups using Botai, AG3, MA1 and West_Siberia_N separately as ancestry proxies. We find ANE-related ancestry appears in centralMongolia_preBA and Fofonovo_EN only to a minor extent; ANE ancestry is not present in eastMongolia_preBA, which is instead characterized by only ANA-related ancestry (Table S5A).

Early Bronze Age

- New genetic groups: Afanaseivo_Mongolia(2), Chemurchek_southAltai(2), Chemurchek_northAltai(2)
- Published genetic groups: Afanaseivo(23), Okunevo_EMBA(19), and Baikal_EBA(5)

Our dataset adds three main genetic groups during the Early Bronze Age: Afanasievo_Mongolia, Chemurchek_southAltai and Chemurchek_northAltai. We group two individuals from Shatar Chuluu site (SHT001, SHT002) into Afanasievo_Mongolia as both are archaeologically classified into the Afanasievo cultural context and genetically indistinguishable from Afanasievo individuals from the Russian Altai-Sayan region (Figure S5A, S5C; Table S5B).

We group two individuals from Yagshiin Huduu site (IAG001, YAG001) into Chemurchek_southAltai as both are archaeologically classified to the Chemurchek cultural context and cluster together on PCA, providing the first genomic investigation of the Chemurchek culture. We observed that Chemurchek_southAltai has the highest genetic affinity to ANE-related groups (e.g., Botai) and secondary affinity to Iranian-related groups (Figures S5A and S5C). We tested Afanasievo as a potential ancestral source given the geographic overlap and similar burial posture between the Afanasievo and Chemurchek cultures (Taylor et al., 2019), however the 2-way model with Afanasievo as one of the two sources fails (Table S5B). The model also fails when using Okunovo (a neighboring group contemporaneous with Chemurchek that succeeds the Afanasievo culture) either as Afanasievo + Okunevo or Okunevo + Iranian (Table S5B). To further investigate the Iranian-related ancestry among the Chemurchek_southAltai, we tested four published groups from the BMAC genetic cluster (Gonur1_BA, Bustan_BA, Dzharkutan1_BA, and Sappali_Tepe_BA), four Chalcolithic/Bronze Age Iranian groups (Hajji_Firuz_C, Tepe_Hissar_C, Seh_Gabi_C, and Shahr_I_Sokhta_BA1), Eneolithic Turkmenistan and Tajikistan groups (Paikhai_EN, Sarazm_EN, and Tepe_Anau_EN) and Mesolithic Caucasus Hunter-Gatherer (CHG). Interestingly, 2-way models consisting of Botai + BMAC genetic cluster groups adequately model Chemurchek_southAltai with ~40% ancestry proportion from the latter, while preceding Eneolithic Turkmenistan/Tajikistan groups do not (Table S5B). Spatiotemporally more distant groups, such as Chalcolithic Iranian groups or CHG, also adequately model Chemuchek_southAltai with similar ancestry proportions (Table S5B). The 3-way model consisting of Botai + BMAC + Afanasievo returns a positive contribution from Afanasievo but it is not significantly different from zero (11 ± 7%; Table S5B). Thus, despite some cultural similarities between steppe groups (Afanasievo, Okunevo) and Chemurchek, we observe a negligible level of genetic influence from the steppe populations among the Chemurchek_southAltai individuals analyzed here.

We find that Chemurchek_southAltai has a close genetic affinity to Dali_EBA (Figure S5A), an individual dating to ca. 2650 BCE with poor burial context from southeastern Kazakhstan who has admixed ANE-Iranian ancestry (see Narasimhan et al., 2019). Applying the same 2-way admixture models using Dali_EBA for comparison, we found that Dali_EBA also requires an additional Iranian-related ancestry but in a smaller proportion and that the models with Afanasievo as a source do not fit, replicating the findings in Chemurchek_southAltai.

The Chemurchek_northAltai genetic cluster, consisting of two female Chemurchek individuals (KUM001, KUR001) from the northern Altai, shows high genetic affinity to ANA groups, with a small proportion of ancestry consistent with Chemurchek_southAltai (Figures S5A, S5C; Table S5B). Like other Chemurchek or Chemurchek-like burials in the northern Altai, their mortuary architecture lacks some features that are classically associated with burials further south.

Middle and Late Bronze Age

- New genetic groups: Altai_MLBA(7), Ulaanzuukh_SlabGrave(11/16), UAA001(1), KHI001(1), UUS001(1), KHU001(1) and TSI001(1)
- Published genetic groups: Khövsgöl_LBA(17), ARS017(1), ARS026(1), Sintashta_MLBA(37), Krasnoyarsk_MLBA(18)

Our dataset adds two main genetic groups in the Eastern Steppe during the MLBA - Altai_MLBA and Ulaanzuukh_SlabGrave - to the previously published Khövsgöl_LBA from northern Mongolia (Jeong et al., 2018). Our new data substantially expand the geographic scope of genetically characterized MLBA populations in Mongolia, and reveal an overall picture of the population structure of the MLBA Eastern Steppe. The Altai_MLBA group contains seven individuals from the Altai-Sayan region (BER002, BIL001, ULZ001, ARS026, SBG001, ULI001, ULI003), who are admixed between the Western Steppe gene pool associated with Srubnaya/Sintashta/Andronovo cultures ("steppe_MLBA") and the one associated with Khövsgöl_LBA/Baikal_EBA. Although the ancestry proportion estimates within this group vary along a cline, the Altai_MLBA represents the formation of a gene pool incorporating a substantial genetic influx from Western Steppe herders. Thus we classified them into one genetic group despite their archaeological and cultural differences (DSKC and unclassified burial types). This also explains the genetic profile of one outlier from Khövsgöl_LBA (ARS026), who now genetically falls within the Altai_MLBA group. Of note, one member of this group, ULZ001, is found not in the Altai, but in far eastern Mongolia.

The other genetic cluster, Ulaanzuukh_SlabGrave, contains 11 individuals with Ulaanzuukh burial type (BUL001, BUL002, ULN001, ULN002, ULN003, ULN005, ULN006, ULN007, ULN009, ULN010, ULN015) and 5 individuals with Slab Grave burials (see below), from eastern Mongolia. They all are classified into one single genetic group given their strong genetic homogeneity with ANA (Table S5C) and the geographic links between the two. This clustering of Ulaanzuukh and Slab Grave confirms previous archaeological hypotheses that the Slab Grave culture likely emerged out of the Ulaanzuukh gene pool. This genetic cluster also explains another Khövsgöl_LBA outlier, ARS017, who now genetically falls within the Ulaanzuukh_SlabGrave group, as well as a single individual with unknown burial type from central Mongolia, TSI001, who also falls into this cluster. Of note, one male Mönkhkhairkhan individual (KHU001) also has a large proportion of ancestry from Ulaanzuukh_SlabGrave in addition to his main genetic component from Baikal_EBA (Table S5C). Together, the individuals ARS017, TSI001, and KHU001 suggest contact with the Ulaanzuukh_SlabGrave group in northern, central Mongolia, even though these individuals were buried according to local burial customs. Overall, this Ulaanzuukh_SlabGrave genetic cluster is a continuation of the ANA easternMongolia_preBA gene pool (represented by SOU001) of 3,000 years earlier.

We also identified three outliers which do not fall into any of the three genetic clusters described above. UAA001 (Mönkhkhairkhan) from the Altai is well-fitted with 3-way admixture model using Afanasievo, Baikal_EBA and Gonur1_BA (Table S5C), despite the fact that they date to ~1500 years after the Afanasievo culture. KHI001 (unclassified culture) from the Altai, is well-fitted with 3-way admixture model using Sintatash, Baikal_EBA and Gonur1_BA (*p-value* = 0.056; Table S5C), presenting minor genetic component from Gonur1_BA. Alternatively, KHI001 can also be modeled as a 2-way admixture between Afanasievo and Khövsgöl_LBA (*p-value* = 0.117; Table S5C); however, this model has a lower priority than the former model. UUS001 (DSKC) from Khövsgöl province is well-fitted with 3-way model using Sintashta, eastMongolia_preBA and Gonur1_BA (Table S5C). Given the temporal discordance between UUS001 and the eastMongolia_preBA individual (~3,000 years), it is more likely that the admixing partner for UUS001 was related to the Ulaanzuukh cluster; Ulaanzuukh shares a high degree of ancestry with eastMongolia_preBA and is contemporaneous with the UUS001 individual, and some Ulaanzuukh individuals plot very close to the eastMongolia_preBA individual - SOU001 in PCA.

Early Iron Age

- New genetic groups: Chandman_IA(9), Ulaanzuukh_SlabGrave(5/16),
- Published genetic groups: Tagar(8), CentralSaka(6), TianShanSaka(10), Kazakhstan_Berel_IA(2; Pazyryk culture)

Our dataset adds two main genetic groups during EIA: one represented by Ulaanzuukh_SlabGrave and the other represented by the site of Chandman Mountain associated with the Sagly/Uyuk culture (Chandman_IA). In addition to the 11 Ulaanzuukh burials described above, four Slab Grave individuals (BOR001, DAR001, MIT001, SHU001) from eastern Mongolia also presented a homogeneous genetic profile with Ulaanzuuk and thus were merged into the Ulaanzuukh_SlabGrave analysis group (Table S5C). Interestingly, PTO001, a Trans-Baikal individual who is also archaeologically classified as Slab Grave, has a genetic profile that matches other Slab Grave individuals from eastern Mongolia, and we also merged PTO001 into the Ulaanzuukh_SlabGrave genetic cluster. The genetic profile of PTO001 is consistent with an archaeologically described expansion of the Slab Grave culture into the Baikal region during EIA (Losey et al., 2017).

The contemporaneous Chandman_IA from the Altai-Sayan region in western Mongolia has a genetic profile that matches the preceding Altai_MLBA cline. Since all individuals are from a single site and cluster together on PCA, we group them into a single analysis

unit (“Chandman_IA”). Here, we use the Andronovo-associated dataset Krasnoyarsk_MLBA as the representative central steppe_MLBA group for admixture modeling because it is geographically closest to our test EIA groups. We first tested a 2-way admixture model of Krasnoyarsk_MLBA + Baikal_EBA, but it failed to adequately model the Chandmand_IA cluster, as did Karsnoyarsk_MLBA + Khövsgöl_LBA. Further changing the steppe_MLBA source from Karsnoyarsk_MLBA to Sintashta_MLBA did not rescue the 2-way admixture model. We then attempted a 3-way admixture model by adding Iranian-related ancestry as the third source, using a BMAC group from the Gonur Tepe site (Gonur1_BA) as a proxy. Using Krasnoyarsk_MLBA as the Steppe proxy, we observed 51.3% of Steppe, 42.2% of Baikal_EBA and 6.5% of Iranian ancestry in Chandman_IA (Table S5D).

Because it is *a priori* quite unlikely due to a long-distance migration from Bactria/Iran specific to Chandman_IA, we next applied the same 3-way models of Krasnoyarsk_MLBA/Sintashta_MLBA+Baikal_EBA+Gonur1_BA to four Iron Age central Asian groups (Tagar from Minusinsk Basin, Central Saka from central Kazakhstan, Kazakhstan_Berel_IA from eastern Kazakhstan, and Tian Shan Saka from Kyrgyzstan) and also to the Final Bronze Age group Karasuk. We observed that Iranian-related ancestry proportions range from ~7%–28% in the tested Iron Age groups, while not required for Karasuk. In particular, the Tian Shan Saka, geographically closest to the Gonur Tepe site, has the highest amount of estimated Iranian-related ancestry. Because of cultural connections between the Sagly/Uyuk of Chandman_IA and the Saka generally (see section 2.4 above), it is possible that Saka and related groups in Tian Sian, Fergana and Transoxiana/Turan (such as the sampled Tian Shan Saka) are the proximal source of the Iranian ancestry in the Iron Age groups further to the north, such as Chandman_IA. To narrow down the spatiotemporal origin of this Iranian-related ancestry, we tested 3-way models using alternative Iranian-related groups as the proxy in the Tian Shan Saka: (1) three other post-BMAC groups (Bustan_BA, Dzharkutan1_BA, and Sappali_Tepe_BA) that fall into the BMAC genetic cluster with Gonur1_BA (Narasimhan et al., 2019), (2) Shahr_I_Sokhta_BA1 from the southeastern corner of Iran, (3) three Chalcolithic Iranian groups (Hajji_Firuz_C, Tepe_Hissar_C, Seh_Gabi_C), (4) two Iron Age groups from Pakistan (Katelai_IA, Loebanr_IA), (5) Eneolithic groups from Turkmenistan such as Geoksyur_EN, Parkhai_EN and Tepe_Anau_EN, and (6) Sarazm_EN fromwestern Tajikistan. All Iranian-ancestry proxies mentioned above except Hajji_Firuz_C and Seh_Gabi_C from the Zagros provide a well-fitted 3-way model (Table S5E). Therefore, for the Iron Age Eastern Steppe, genetic data alone can only narrow down the source of the Iranian ancestry to a broad region east of the Caspian Sea. Taken in context, though, we propose that this ancestry likely arrived via a local contact around the Transoxiana/Sogdiana region (i.e., the border between Kazakhstan, Uzbekistan and Kyrgyzstan).

For the prehistoric genetic groups described above, we used DATES to estimate the date of admixture between Western ancestry sources (WSH or the Iranian-related groups) and local ancestry sources (i.e., Khovsgol_LBA or Baikal_EBA) (Figure S6). As shown in Figure S6A, the estimated admixture date between Sintashta and Baikal_EBA for the Karasuk and Tagar is consistent with the admixture date observed in Altai_MLBA - at around 3,500 BP. For the Central Saka, Pazyryk (Kazakstan_Berel_IA) and Sagly/Uyuk (Chandman_IA), the admixture date is estimated to be a few centuries later, and the most recent admixture date is estimated for the Saka from Tian Shan. Notably, we find that the estimated admixture dates between Gonur1_BA and Baikal_EBA in the Iron Age groups are roughly consistent with the admixture dates for Sintashta (Figure S6A). However, because we are using a method designed for dating a 2-way admixture on what is best modeled as 3-way admixture in our study, we caution that these admixture dates should be interpreted with care.

Xiongnu Empire

• New genetic groups: earlyXiongnu_west(6), earlyXiongnu_rest(6), SKT007(1), lateXiongnu(24), lateXiongnu_sarmatian(13), lateXiongu_han(8), TAK001(1), TUK002(1)

• Published genetic groups: Xiongnu_WE(2), Xiongnu_royal(1, DA39.SG), Han_2000BP(2)

Our dataset reveals a great deal of previously uncharacterized genetic diversity during the Xiongnu period. For individual modeling, we tested every possible combination of five main ancestries: Steppe (Krasnoyarsk_MLBA, Sintashta, Srubnaya, Sarmatian, Chandman_IA), Gonur1_BA, Khövsgöl_LBA, Ulaanzuukh_SlabGrave, and Han. Considering the low resolution of individual modeling, we report selected working models that work for many individuals belonging to the same time period and archaeological context and that reflect qualitative trends observed in PCA. We observed that Iron Age Chandman_IA is a good Steppe ancestry proxy for many Xiongnu individuals, but there are also many who have western Eurasian ancestry in higher proportion than that of Chandman_IA. These individuals with high western Eurasian ancestry proportion show strong affinity to the Iranian-related ancestry that cannot be explained by the earlier Late Bronze Age steppe groups (e.g., Krasnoyarsk_MLBA, Sintashta_MLBA or Srubnaya). Instead, Gonur1_BA or Iron Age Sarmatian fit better with the genetic profile required. Also, a few individuals fall into the eastern Eurasian cline along PC2 and are explained as a combination of the eastern Eurasian gene pools, Ulaanzuukh_SlabGrave and present-day Han Chinese, without contribution from western Eurasian sources (Table S5F). We used high-coverage whole genome sequences of present-day Han Chinese (“Han.DG”; n = 4) as a proxy for the ancestry component that is currently broadly distributed across northern China and distinct from the component represented by Ulaanzuukh_SlabGrave further to the north. This is to achieve statistical power in our admixture modeling given that there are to date very few available ancient genomes that reflect this ancestry component. This is due to the fact that ancient China, Korea, Japan, and Southeast Asia remain mostly unsampled. We fully acknowledge the genetic diversity present within contemporary Han Chinese populations, and do not intend to claim by our admixture modeling a specific connection between the ancient populations within our study and present-day ethno-cultural identities.

For the group-based qpAdm modeling, we split Xiongnu into two categories based on their age - early Xiongnu and late Xiongnu. We further split early Xiongnu into two subgroups, earlyXiongnu_west (SKT010, SKT001, SKT003, SKT009, SKT008, AST001) and earlyXiongnu_rest (JAG001, SKT002, SKT004, SKT005, SKT006, SKT012), based on their individual modeling results, leaving out one individual outlier - SKT007 (Khövsgöl_LBA-like). The two previously published Xiongnu individuals grouped as “Xiongnu_WE”

show a similar genetic profile to earlyXiongnu_rest, are dated to the early Xiongnu period, and are from the same valley as the two early Xiongnu sites (SKT and AST) in our dataset (Table S5F). For the late Xiongnu, we summarized their individual modeling results in Table S5G. Based on the individual modeling results, we set up three subgroups within late Xiongnu individuals to highlight key demographic processes and to use them for specific analyses such as sex-biased gene flow. First, we assigned 24 of 47 individuals into the main lateXiongnu group (BTO001, CHN010, DEL001, DOL001, IMA001-IMA008, JAA001, KHO006, KHO00, SAN001, SOL001, TEV002, TEV003, TUK003, UGU004, UGU011, ULN004, UVG001; Table S5F-G); this group is well modeled as a mixture of two main Iron Age clusters, Chandman_IA+Ulaanzuukh_SlabGrave ($p = 0.316$; 76.6 ± 0.8% from Ulaanzuukh_SlabGrave). Another 13 individuals have more western Eurasian ancestry than Chandman_IA and thus require a different western Eurasian source. Two of them (NAI002, BUR001) are explained by Chandman_IA+Gonur1_BA, a model for earlyXiongnu_west, but the remaining 11 need Sarmatian contribution, including three that are cladal to Sarmatian (BUR003, TM001, UGU010). Taken all 13 individuals as a group (lateXiongnu_sarmatian; BRL002, BUR001-BUR004, DUU001, HUD001, NAI001, NAI002, TMI001, UGU005, UGU006, UGU010), we infer a major contribution from a Sarmatian-related source into this group (75.7 ± 2.8%; Table S5F-G). On the other hand, we grouped eight individuals (ATS001, BAM001, BRU001, EME002, SON001, TUH001, TUH002, YUR001) into the third group lateXiongnu_han based on their affinity to Han Chinese and other East Asian populations that Ulaanzuukh_SlabGrave cannot explain (37.2 ± 10.6% from Han.DG; Table S5F-G). The previously published Xiongnu_royal individual shows substantial Han-related ancestry (Table S5F), similar to our lateXiongnu_han group. Further, the late Xiongnu individual YUR001 is an extreme East Asian outlier, who genetically resembles "Han_2000BP," two Han empire soldiers recovered from a mass grave near a Han fortress in the southern Gobi (Damgaard et al., 2018b). These two groups, lateXiongnu_sarmatian and lateXiongnu_han, robustly support influxes of new ancestries both from the west and the east that were not previously observed in early Xiongnu or earlier populations. We left two individuals out of grouping, due to their unusual ancestry profiles: TAK001 mostly resembles Khövsgöl_LBA, and TUK002 is modeled as Chandman_IA+Ulaanzuukh_SlabGrave_Gonur1_BA (Table S5G). In contrast to the strong east-west genetic division among Bronze Age Eastern Steppe populations through the end of the Early Iron Age, the Xiongnu period is characterized by an extreme degree of genetic diversity and heterogeneity that does not have any obvious geographic correlation (Figure S7A).

Early Medieval

• New genetic groups: TUK001(1), earlyMed_Türk(7), TUM001(1), earlyMed_Uyghur(12), OLN007(1)

Our dataset adds two main genetic groups during the early Medieval period in Mongolia: earlyMed_Türk and earlyMed_Uyghur. For each individual, we tested every possible combination of four main ancestries: Steppe (Sarmatian, Alan), Gonur1_BA, Ulaanzuukh_SlabGrave, and Han. The genetic contribution from Iranian-related ancestry becomes even more prominent in Türkic and Uyghur individuals, as seen from well-fitted models using the Alan, an Iranian pastoral population from the Caucasus (Table S5H). Overall, the Türkic and Uyghur individuals in this study show a high degree of genetic diversity, as seen in their wide scatter across PC1 in Figure 2. TUK001(250-383 CE), the earliest early Medieval individual in our dataset from a Xiongnu site with a post-Xiongnu occupation, has the highest western Eurasian affinity. This individual is distinct from Sarmatians, and likely to be admixed between Sarmatians and populations with BMAC/Iranian-related ancestry (Table S5H). Among the Türkic period individuals, TUM001 is a genetic outlier with mostly East Asian (Han_2000BP-like) ancestry. This individual was buried together with a knife and two dogs within the ramp of a Türkic era mausoleum. The mausoleum's stone epitaph indicates that it was constructed for a diplomatic emissary of the Pugu tribe who was allegiant to the Chinese Tang Empire. His cremated remains were found within the tomb; TUM001 was likely this emissary's servant (Ochir et al., 2013).

With respect to Uyghur burials, many consist of collective graves, and it has been suggested that such graves may contain the remains of kin groups (Erdenebat, 2016). We examined one such collective grave (grave 19) at the site of Olon Dov; however, of the six individuals analyzed in grave 19, there were no first degree relatives (parent-offspring pairs or sibling), and only two individuals (OLN002 and OLN003) exhibited a second degree (avuncular, grandparent-grandchild, or half-sibling) relationship. One Uyghur individual (OLN007) had markedly higher proportions of Han-related East Asian ancestry that cannot be explained by Ulaanzuukh_SlabGrave, and therefore grouped separately from the other earlyMed_Uyghur individuals (Table S5H).

Late Medieval

• New genetic groups: lateMed_Khitan(3), lateMed_Mongol(61), SHU002(1)

Our dataset adds two main genetic groups during late Medieval in Mongolia: lateMed_Khitan and lateMed_Mongol. We used the same modeling strategy as used for the early Medieval period, and additionally explored the genetic cladality between every individual from the Mongol period and from modern Mongolic-speaking populations via qpWave (Figure S7B; Table S5J). Relatively few Khitan individuals ($n = 3$) were available for analysis, but all show high ANA-related ancestry (Table S5I). Mongol-era individuals ($n = 61$) are genetically more diverse and are cladal with modern Mongolic-speaking populations (Figure S7B). SHU002 is a single individual dated to the late Medieval period however without recognizable Mongol-like burial feature. Overall, Mongol period individuals characterized by a remarkable decrease in Western Eurasian ancestry compared to the preceding 1,600 years. They are best modeled as a mixture of ANA-like and East Asian-like ancestry sources, with only minor Western genetic ancestry. In addition, nearly a third of historic Mongol males (12/38) have Y haplogroup C2b, which is also widespread among modern Mongolians (Figure S3; Table S6); C2b is the presumed patrilineage of Genghis Khan (Zerjal et al., 2003).

7. Discussion

7.1 MSMC-IM and outlook

This thesis demonstrates how new analytical methods and ancient DNA can help us in understanding population structure and human demographic history. Sequencing technology innovation has given rise to an exponentially increasing amount of genomic data. This scale of both modern or ancient genomic data relies on novel analytical approaches to interpret this type of information. Given the emerging availability of high-coverage genomic data, methods using linkage disequilibrium from full genome sequences can infer population size changes (e.g. PSMC, MSMC, SMC++) (Li et al., 2009; Schiffels & Durbin, 2014b; Terhorst et al., 2017) and population separation process as a function of time (e.g. MSMC) (Schiffels & Durbin, 2014a).

Manuscript A of the thesis introduced a new method built on the MSMC framework to investigate pairwise population separation in more complex scenarios, and shed new insights into the deep population structure in African populations in particular. MSMC-IM takes within- and cross-population coalescent rates estimated from MSMC2 as input, and then estimates time-dependent migration rates and effective population sizes for a pair of populations. With this new method, we use a more gradual concept of population separation in a continuous IM model, discarding the stringent definition of split time and short-term migration rate in a classic IM model (Hey, 2010). In this continuous IM model, only time-dependent migration rates can define the process of population separation and gene flow, which allows a wider time frame when inferring these demographic events. Thus the prominent advantages of MSMC-IM are - i) the flexibility of the demographic model and ii) the continuity of time-dependent migrant rates ranging from present-day to million years ago. It reveals extremely deep population structure in African populations, such as in the ancestors of San and Mbuti, tracing back to over 1 million years ago.

Previous studies (Knight et al., 2003; Pickrell et al., 2012b; Schlebusch et al., 2012; Schlebusch & Jakobsson, 2018; Tishkoff et al., 2007) have identified the southern African hunter-gatherer population i.e. the San, harbours the deepest branches of modern humans in the history of population diversification, dated to approximately 300 thousand years ago (kya).
But the extremely deep signal we detected in San and Mbuti, appears 1 million years ago, far beyond the deepest split time among modern human populations. Although accumulated evidence reveals the legacy of Neaderthal introgression in various African populations (Gurdasani et al., 2015; Prüfer et al., 2014a; S. Wang et al., 2013), the estimated split time between Neadertal and modern humans (~450 kya (Prüfer et al., 2014a)) is not deep enough for explaining such a deep structure. Such deep time depth of divergence has been previously estimated for the hypothesized “super-archaic” population who admixed into Denisovans (between 1.1 and 4 million years ago) (Prüfer et al., 2014b).

Thus this deep structure more likely suggests that the introgression from unknown archaic hominins diverged earlier from the main human lineages, rather than any main separation processes involving San or Mbuti. The hypothesis of unknown archaic introgression in Africa has been increasingly mentioned in recent studies, focusing on western African pygmies like Yoruba and Mandenka (Durvasula & Sankararaman, 2018; Hammer et al., 2011; Hsieh et

al., 2016; Lachance et al., 2012; Lorente-Galdos et al., 2019; Plagnol & Wall, 2006; Xu et al., 2017). Our study reveals the remarkable time depth of the hypothesized introgression from an unknown archaic population in Africa, particularly in indigeneous African hunter-gatherers San and Mbuti.

Furthermore, MSMC-IM detects the complex pattern of post-split migration and archaic introgression in Africa as a more continuous process rather than a single pulse admixture event. The long-lasting signal of non-zero migration rates among African populations, is more likely to have resulted from repeated isolation and admixture of two or more archaic populations that co-existed for a long time, in which case we do not observe a single separation between African and an unknown archaic population, but a mosaic period of merged different separation processes across archaic populations and modern Africans. This repeated isolation and admixture appears to have happened often in the past in Africa. Manuscript B of this thesis reports the complex admixture process between hunter-gatherers and pastoralists in the last 3,000 years in eastern Africa, which I will detail in the second part of the Discussion.

Despite the continuous character of MSMC-IM enabling reconstruction on the dynamics of population separation across worldwide populations, it can not be applied to multiple populations at a time. Only by cross-comparing pairwise results can researchers grasp an overall picture of multi-population separation. An important direction of future work is to develop multi-population demographic concepts such as graph-based models, although it might be technically infeasible to extend the current concept of two island-like populations defined by continuous-time migration rates only to multiple populations. To date, widely used graph-models for reconstructing phylogenetic trees include G-Phocs (Gronau et al., 2011), momi2 (J. Kamm et al., 2019), rarecoal (Schiffels et al., 2016), TreeMix (Pickrell & Pritchard, 2012), qpGraph (Patterson et al., 2012). An important feature of these methods is that they are based on the allele frequency spectrum information (except for G-Phocs). TreeMix and qpGraph aims to build phylogenetic trees via modelling population splits in bifurcating order, and adding inferred gene flow across branches. Both can be applied to multiple populations to reconstruct the topology structure of demographic events, though without precise estimates on split time. While approaches like rarecoal, momi2 and G-Phocs can make inferences in exact numbers on key demographic parameters like divergence time, effective population size, and migration rate, these occur under a strictly defined demographic model that reduces the topology flexibility when building the phylogenetic tree. A major advantage of these allele frequency based methods is that it can be scaled to an arbitrary number of individuals without much computational burden. And therefore can be applied to multiple populations. For instance, SMC++ combined this advantage with the SMC framework, so that it can jointly infer split times and effective population size for hundreds of unphased genomes (Terhorst et al., 2017). But, in the SMC++ approach, a split time is arbitrarily defined and designed for pairwise runs, while the phylogenetic structure of multiple populations is missing.

7.2 aDNA - challenges and future

The field of human archaeogenetics has been growing rapidly in the last decade. Ancient DNA studies have added great resolution to the pre- and historic pictures of population movements, interactions and transformations across all inhabited continents. However, the continent harboring the most genetic diversity - Africa, and the most populous continent - Asia, has been seriously underrepresented in the current dataset of available ancient genomic data. I analyzed a significant number of ancient genomes in Manuscript B and C of this thesis to address some major prehistoric questions in Africa and Asia.

The fundamental challenge in aDNA studies is preservation, which is the determining factor of whether we can achieve usable amounts of DNA for genomic analyses. In general, the environmental preservation conditions and the age of samples are critical for yielding aDNA. The tropical environment in Africa makes it even more challenging for aDNA retrieval as DNA molecules degrade faster when the environment is hot, humid and acidic (Lindahl, 1993). In Manuscript B of this thesis, we screened 57 skeleton materials from sub-Saharan Africa and successfully obtained aDNA for only 23 materials (belong to 20 individuals). Though SNP capture sequencing was used in this study which usually guarantees a higher success rate than shotgun sequencing, we obtained only 40% of success rate. In another African aDNA study using SNP capture sequencing, 43 out of 77 skeleton elements yielded successful aDNA, i.e. 55% success rate (Prendergast et al., 2019). Nowadays it is widely accepted that the *pars-petrosa* of the temporal bone and the chamber of teeth has better preservation for aDNA (P. B. Damgaard et al., 2015; Hansen et al., 2017; Pinhasi et al., 2015), which is the sampling strategy we used in both manuscript B and C. In contrast to African samples, the samples from the Eastern Steppe had surprisingly good aDNA preservation - 214 out of 246 skeleton materials yielded analysable genomic data. This might be attributed to the unique environmental condition - dry and cold weather in Mongolia (Allentoft et al., 2012).

Another challenge in aDNA studies is sampling bias, including both geographic and temporal bias. Undoubtedly, the ideal way of conceiving an aDNA study is to perform sampling based on designed research questions and hypotheses which require preliminary knowledge from either genomic or archeological studies. But it is often challenging given the scarcity or the availability issue of skeleton materials in some regions. For instance, currently available ancient African genomes are centered on eastern Africa and southern Africa (Prendergast et al., 2019; Schlebusch et al., 2017; Skoglund et al., 2017), while in contrast there is a clear lack of ancient genomic data from western, central and northeastern Africa. To understand the spread of pastoralists from northeast Africa to eastern Africa, ancient individuals from northeast Africa (directly dated to Pastoral Neolithic or Pastoral Iron Age), would be particularly useful. To reconstruct the migration history of Bantu-speaking populations, we would need ancient genomic data along the hypothesised dispersal routes between western Africa and central Africa or between western Africa and southern Africa (Crowther et al., 2018). Samples from Botswana in manuscript B provides a good example of resolving archaeological question on the arrival order of pastoralists and Bantu-related farmers in southern Africa, where three individuals from the first millennium in the Okavango Delta (the hypothesized first arrival spot of pastoralism in southern Africa (Cordova, 2018)) demonstrate the earlier arrival of pastoralists than farmers. Considering the unique diversity of genetics, languages and ethnicities in Africa, we call for more efforts on cross-disciplinary

collaborations in the future to detangle African prehistory through the inextricable correlations among the different disciplines.

A good example of thorough sampling across various regions and a long time transect, is manuscript C of this thesis. The newly reported dataset in manuscript C records the dynamics of population history spanning from the mid-Holocene to the late Medieval, including historical pastoral empires. We found a shared mid-Holocene gene pool stretching from Lake Baikal to the Far East by analyzing two individuals from central and eastern Mongolia, together with individuals from Lake Baikal and Devils Cave in the Far East (Sikora et al., 2019). More samples from Tuva, Altai-Sayan and northeastern China would enlarge the geographic zone of this shared gene pool to a broader scale. Entering the Bronze Age, Yamnaya/Afanasievo steppe herders expanded to the center of Mongolia but did not leave long-lasting genetic signals, given current data, unlike in Europe (Allentoft et al., 2015; Haak et al., 2015; Mathieson et al., 2015). Subsequent Chemurchek pastoralists, represented by two individuals from one site in the Altai, derive their ancestry from ANE-related pastoralists' migration and Iranian-related ancestry, probably tracing back through Xinjiang or mountainous Central Asia where Chemurchek was once widely spread (Jia & Betts, 2010). aDNA from Chemurchek burials in Xinjiang or Central Asia would help in improving understanding of the signal of Iranian-related ancestry. During the Middle Late Bronze Age, we observed a strong spatial-genetic correlation forming a tripartite genetic structure in northern, western and eastern Mongolia correspondingly. A new migration wave of Sintashta steppe herders came in and admixed with local communities to different degrees in the northern and western Mongolia, while the eastern area maintains the previous mid-Holocene gene pool. Geographic barriers and cultural impact might have been contributing factors to such spatial-genetic structure.

As human genetic diversity aligns well with geography in the past as well as present (Novembre et al., 2008; Peter et al., 2019), we are now in need for more new tools on modelling and visualising such continental-level population movements and settlements. Application of equilibrium stepping stone models, coupled with Fst (a measure of genetic distance), is a good practice for modelling spatial population structure (Kimura & Weiss, 1964; Slatkin, 1991). For example, EEMS visualizes spatial population structure using estimated effective migration rates on a geographic map employing an approximation of a general stepping-model model (Petkova et al., 2016). Nevertheless, previous studies and Manuscript C of this thesis showed that a spatial-genetic correlation does not always follow the "isolation by distance" rule (i.e. genetic similarity decays with geographic distances), which theoretically is only valid in equilibrium populations of constant allelic frequencies (Kimura & Weiss, 1964; Peter & Slatkin, 2013; Ramachandran et al., 2005). Population migration is more likely to be a temporally gradual process resulting in fluctuating allelic frequencies. Thus it is of equal importance to visualise temporal population structure for understanding human demography. Expanding aDNA dataset makes it possible to track human mobility changes through time. For instance, Loog et al proposed the S_{max} statistic for modelling spatiotemporal structure on hundreds of published ancient west Eurasians (Loog et al., 2017).

Since thousands of ancient genomes are being generated, aDNA field is from various perspectives, in the middle of a transition. On one hand, the current aDNA dataset is biased towards prehistoric periods focusing on the early peopling history at continental levels. Only

limited publications (Antonio et al., 2019; Schiffels et al., 2016; Vai et al., 2019; Veeramah et al., 2018) focus on Iron Age and Medieval eras when migration of tribes and formation of empires are known in historical records. In future studies of historical periods, it would be interesting to investigate whether genetic results are consistent with written records, just like investigating whether gene flow exchanges between ancient groups are in line with interactions between cultures. For example, manuscript C of this thesis reveals genetic origins of the Xiongnu empire and Mongol empire, which both are well known for the history of conquering lands across the Eurasian continent; where the former is genetically very heterogeneous - consistent with their high mobility - and the latter in contrast is more homogenized shifting, to East Eurasian ancestry. Furthermore, clear information of archeological cultures and historical documents make it easier to raise clearly defined genetic research questions, when designing an aDNA study. Overall we call for more aDNA studies on Iron Age and Medieval periods in the future.

On the other hand, we need more new summary statistics on tackling other key questions in human evolution, such as adaptation of complex traits and selection in response to the environment or subsistence changes. For example, pastoralism is still prevalent in present-day Africa and the Eastern Steppe, however in manuscript B and C we do not find positive genetic evidence for lactase persistence (LP) in ancient individuals who are identified as linked to dairy pastoralism. One possibility is that the LP trait in the past is associated with allelic mutations different from modern-day herders. Ideally, we could perform genome-wide selection scans for novel LP-related loci with a big aDNA dataset including only individuals practicing dairy pastoralism across various periods and regions. A previous large-scale aDNA study in Europe revealed selection at multiple loci related to diet, which may have been linked to subsistence changes and population movement, and also some loci related to pigmentation, immunity and height (Mathieson et al., 2015). The study of the evolution of complex traits like immunity, height or any other polygenic traits, would benefit enormously from the expanding aDNA dataset.

8. References

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.

Allentoft, M. E., Collins, M., Harker, D., Haile, J., Oskam, C. L., Hale, M. L., Campos, P. F., Samaniego, J. A., Gilbert, M. T. P., Willerslev, E., Zhang, G., Scofield, R. P., Holdaway, R. N., & Bunce, M. (2012). The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings. Biological Sciences / The Royal Society*, *279*(1748), 4724–4733.

Allentoft, M. E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P. B., Schroeder, H., Ahlström, T., Vinner, L., Malaspinas, A.-S., Margaryan, A., Higham, T., Chivall, D., Lynnerup, N., Harvig, L., Baron, J., Della Casa, P., Dąbrowski, P., … Willerslev, E. (2015). Population genomics of Bronze Age Eurasia. *Nature*, *522*(7555), 167–172.

Anthony, D. W. (2010). *The Horse, the Wheel, and Language: How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World*. Princeton University Press.

Antonio, M. L., Gao, Z., Moots, H. M., Lucci, M., Candilio, F., Sawyer, S., Oberreiter, V., Calderon, D., Devitofranceschi, K., Aikens, R. C., Aneli, S., Bartoli, F., Bedini, A., Cheronet, O., Cotter, D. J., Fernandes, D. M., Gasperetti, G., Grifoni, R., Guidi, A., … Pritchard, J. K. (2019). Ancient Rome: A genetic crossroads of Europe and the Mediterranean. *Science*, *366*(6466), 708–714.

Briggs, A. W., Good, J. M., Green, R. E., Krause, J., Maricic, T., Stenzel, U., Lalueza-Fox, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Schmitz, R., Doronichev, V. B., Golovanova, L. V., de la Rasilla, M., Fortea, J., Rosas, A., & Pääbo, S. (2009). Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science*, *325*(5938), 318–321.

Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S., & Akey, J. M. (2018). Analysis of

Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell*, *173*(1), 53–61.e9.

Burbano, H. A., Hodges, E., Green, R. E., Briggs, A. W., Krause, J., Meyer, M., Good, J. M., Maricic, T., Johnson, P. L. F., Xuan, Z., Rooks, M., Bhattacharjee, A., Brizuela, L., Albert, F. W., de la Rasilla, M., Fortea, J., Rosas, A., Lachmann, M., Hannon, G. J., & Pääbo, S. (2010). Targeted investigation of the Neandertal genome by array-based sequence capture. *Science*, *328*(5979), 723–725.

Cooper, A., & Poinar, H. N. (2000). Ancient DNA: do it right or not at all [Review of *Ancient DNA: do it right or not at all*]. *Science*, *289*(5482), 1139.

Cordova, C. (2018). *Geoarchaeology: The Human-Environmental Approach*. Bloomsbury Publishing.

Crowther, A., Prendergast, M. E., Fuller, D. Q., & Boivin, N. (2018). Subsistence mosaics, forager-farmer interactions, and the transition to food production in eastern Africa. *Quaternary International: The Journal of the International Union for Quaternary Research*, *489*, 101–120.

Damgaard, P. B., Margaryan, A., Schroeder, H., Orlando, L., Willerslev, E., & Allentoft, M. E. (2015). Improving access to endogenous DNA in ancient bones and teeth. In *Scientific Reports* (Vol. 5, Issue 1). https://doi.org/10.1038/srep11184

Damgaard, P. de B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., Moreno-Mayar, J. V., Pedersen, M. W., Goldberg, A., Usmanova, E., Baimukhanov, N., Loman, V., Hedeager, L., Pedersen, A. G., Nielsen, K., Afanasiev, G., Akmatov, K., Aldashev, A., Alpaslan, A., … Willerslev, E. (2018). 137 ancient human genomes from across the Eurasian steppes. *Nature*, *522*, 207.

de Barros Damgaard, P., Martiniano, R., Kamm, J., Moreno-Mayar, J. V., Kroonen, G., Peyrot, M., Barjamovic, G., Rasmussen, S., Zacho, C., Baimukhanov, N., Zaibert, V., Merz, V., Biddanda, A., Merz, I., Loman, V., Evdokimov, V., Usmanova, E., Hemphill, B., Seguin-Orlando, A., … Willerslev, E. (2018). The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science*, *360*(6396).

https://doi.org/10.1126/science.aar7711

de Filippo Cesare, Bostoen Koen, Stoneking Mark, & Pakendorf Brigitte. (2012). Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1741), 3256–3263.

Durvasula, A., & Sankararaman, S. (2018). Recovering signals of ghost archaic admixture in the genomes of present-day Africans. In *bioRxiv* (p. 285734). https://doi.org/10.1101/285734

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*(10), e1003905.

Fu, Q., Meyer, M., Gao, X., Stenzel, U., Burbano, H. A., Kelso, J., & Pääbo, S. (2013). DNA analysis of an early modern human from Tianyuan Cave, China. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(6), 2223–2227.

Gallego Llorente, M., Jones, E. R., Eriksson, A., Siska, V., Arthur, K. W., Arthur, J. W., Curtis, M. C., Stock, J. T., Coltorti, M., Pieruccini, P., Stretton, S., Brock, F., Higham, T., Park, Y., Hofreiter, M., Bradley, D. G., Bhak, J., Pinhasi, R., & Manica, A. (2015). Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. *Science*, *350*(6262), 820–822.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., … Pääbo, S. (2010). A draft sequence of the Neandertal genome. *Science*, *328*(5979), 710–722.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., & Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, *43*(10), 1031–1034.

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., Choudhury, A., Ritchie, G. R. S., Xue, Y., Asimit, J., Nsubuga, R. N., Young, E. H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., … Sandhu, M. S. (2015). The African Genome Variation Project shapes medical

genetics in Africa. *Nature*, *517*(7534), 327–332.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, *5*(10), e1000695.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., Fu, Q., Mittnik, A., Bánffy, E., Economou, C., Francken, M., Friederich, S., Pena, R. G., Hallgren, F., Khartanovich, V., … Reich, D. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, *522*(7555), 207–211.

Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C., & Wall, J. D. (2011). Genetic evidence for archaic admixture in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(37), 15123–15128.

Hansen, H. B., Damgaard, P. B., Margaryan, A., Stenderup, J., Lynnerup, N., Willerslev, E., & Allentoft, M. E. (2017). Comparing Ancient DNA Preservation in Petrous Bone and Tooth Cementum. *PloS One*, *12*(1), e0170940.

Heine, B., & Nurse, D. (2000). *African Languages: An Introduction*. Cambridge University Press.

Hey, J. (2010). Isolation with migration models for more than two populations. *Molecular Biology and Evolution*, *27*(4), 905–920.

Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A., & Wilson, A. C. (1984). DNA sequences from the quagga, an extinct member of the horse family. *Nature*, *312*(5991), 282–284.

Hobolth, A., Andersen, L. N., & Mailund, T. (2011). On computing the coalescence time density in an isolation-with-migration model with few samples [Review of *On computing the coalescence time density in an isolation-with-migration model with few samples*]. *Genetics*, *187*(4), 1241–1243.

Hollfelder, N., Schlebusch, C. M., Günther, T., Babiker, H., Hassan, H. Y., & Jakobsson, M. (2017). Northeast African genomic variation shaped by the continuity of indigenous

groups and Eurasian migrations. *PLoS Genetics*, *13*(8), e1006976.

Honeychurch, W. (2015). *Inner Asia and the Spatial Politics of Empire: Archaeology, Mobility, and Culture Contact*. Springer, New York, NY.

Hsieh, P., Woerner, A. E., Wall, J. D., Lachance, J., Tishkoff, S. A., Gutenkunst, R. N., & Hammer, M. F. (2016). Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. *Genome Research*, *26*(3), 291–300.

Jeong, C., Wilkin, S., Amgalantugs, T., Bouwman, A. S., Taylor, W. T. T., Hagan, R. W., Bromage, S., Tsolmon, S., Trachsel, C., Grossmann, J., Littleton, J., Makarewicz, C. A., Krigbaum, J., Burri, M., Scott, A., Davaasambuu, G., Wright, J., Irmer, F., Myagmar, E., … Warinner, C. (2018). Bronze Age population dynamics and the rise of dairy pastoralism on the eastern Eurasian steppe. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(48), E11248–E11255.

Jia, P. W. M., & Betts, A. V. G. (2010). A re-analysis of the Qiemu'erqieke (Shamirshak) cemeteries, Xinjiang, China. *Journal of Indo-European Studies*, *38*(3/4), 275.

Kamm, J. A., Terhorst, J., & Song, Y. S. (2017). Efficient computation of the joint sample frequency spectra for multiple populations. *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, *26*(1), 182–194.

Kamm, J., Terhorst, J., Durbin, R., & Song, Y. S. (2019). Efficiently Inferring the Demographic History of Many Populations With Allele Count Data. *Journal of the American Statistical Association*, 1–16.

Kimura, M., & Weiss, G. H. (1964). The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. *Genetics*, *49*(4), 561–576.

Kindstedt, P. S., & Ser-Od, T. (2019). Survival in a Climate of Change: The Origins and Evolution of Nomadic Dairying in Mongolia. *Gastronomica: The Journal of Critical Food Studies*, *19*(3), 20–28.

Knight, A., Underhill, P. A., Mortensen, H. M., Zhivotovsky, L. A., Lin, A. A., Henn, B. M.,

Louis, D., Ruhlen, M., & Mountain, J. L. (2003). African Y chromosome and mtDNA divergence provides insight into the history of click languages. *Current Biology: CB*, *13*(6), 464–473.

Kovalev, A. (2014). Earliest European in the heart of Asia: the Chemurchek cultural phenomenon, vol. 2. *Saint Petersburg: Book Antiqua*.

Krause, J., Briggs, A. W., Kircher, M., Maricic, T., Zwyns, N., Derevianko, A., & Pääbo, S. (2010). A complete mtDNA genome of an early modern human from Kostenki, Russia. *Current Biology: CB*, *20*(3), 231–236.

Krause, J., Fu, Q., Good, J. M., Viola, B., Shunkov, M. V., Derevianko, A. P., & Pääbo, S. (2010). The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature*, *464*(7290), 894–897.

Kuznetsov, P. F. (2006). The emergence of Bronze Age chariots in eastern Europe. *Antiquity*, *80*(309), 638–645.

Lachance, J., Vernot, B., Elbers, C. C., Ferwerda, B., Froment, A., Bodo, J.-M., Lema, G., Fu, W., Nyambo, T. B., Rebbeck, T. R., Zhang, K., Akey, J. M., & Tishkoff, S. A. (2012). Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell*, *150*(3), 457–469.

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D. C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., Connell, S., Stewardson, K., Harney, E., Fu, Q., Gonzalez-Fortes, G., Jones, E. R., Roodenberg, S. A., Lengyel, G., Bocquentin, F., … Reich, D. (2016). Genomic insights into the origin of farming in the ancient Near East. *Nature*, *536*(7617), 419–424.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P. H., Schraiber, J. G., Castellano, S., Lipson, M., Berger, B., Economou, C., Bollongino, R., Fu, Q., Bos, K. I., Nordenfelt, S., Li, H., de Filippo, C., Prüfer, K., … Krause, J. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, *513*(7518), 409–413.

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-

genome sequences. *Nature*, *475*(7357), 493–496.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* , *25*(16), 2078–2079.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, *362*(6422), 709–715.

Loog, L., Mirazón Lahr, M., Kovacevic, M., Manica, A., Eriksson, A., & Thomas, M. G. (2017). Estimating mobility using sparse data: Application to human genetic variation. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(46), 12213–12218.

Lorente-Galdos, B., Lao, O., Serra-Vidal, G., Santpere, G., Kuderna, L. F. K., Arauna, L. R., Fadhlaoui-Zid, K., Pimenoff, V. N., Soodyall, H., Zalloua, P., Marques-Bonet, T., & Comas, D. (2019). Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations. In *Genome Biology* (Vol. 20, Issue 1). https://doi.org/10.1186/s13059-019-1684-5

Malaspinas, A.-S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J. Y., Crawford, J. E., Heupink, T. H., Macholdt, E., Peischl, S., Rasmussen, S., Schiffels, S., Subramanian, S., Wright, J. L., Albrechtsen, A., Barbieri, C., … Willerslev, E. (2016). A genomic history of Aboriginal Australia. *Nature*, *538*(7624), 207–214.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., … Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201–206.

Maricic, T., Whitten, M., & Pääbo, S. (2010). Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PloS One*, *5*(11), e14004.

Marjoram, P., & Wall, J. D. (2006). Fast “coalescent” simulation. *BMC Genetics*, *7*, 16.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., Sirak, K., Gamba, C., Jones, E. R., Llamas, B., Dryomov, S., Pickrell, J., Arsuaga, J. L., de Castro, J. M. B., Carbonell, E., … Reich, D. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, *528*(7583), 499–503.

McVean, G. A. T., & Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1459), 1387–1393.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, *11*(1), 31–46.

Mittnik, A., Massy, K., Knipper, C., Wittenborn, F., Friedrich, R., Pfrengle, S., Burri, M., Carlichi-Witjes, N., Deeg, H., Furtwängler, A., Harbeck, M., von Heyking, K., Kociumaka, C., Kucukkalipci, I., Lindauer, S., Metz, S., Staskiewicz, A., Thiel, A., Wahl, J., … Krause, J. (2019). Kinship-based social inequality in Bronze Age Europe. *Science*, *366*(6466), 731–734.

Narasimhan, V. M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., Lazaridis, I., Nakatsuka, N., Olalde, I., Lipson, M., Kim, A. M., Olivieri, L. M., Coppa, A., Vidale, M., Mallory, J., Moiseyev, V., Kitov, E., Monge, J., Adamski, N., … Reich, D. (2019). The formation of human populations in South and Central Asia. *Science*, *365*(6457), eaat7487.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., & Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, *456*(7218), 98–101.

Pääbo, S. (1985). Molecular cloning of Ancient Egyptian mummy DNA. *Nature*, *314*(6012), 644–645.

Pääbo, S. (1989). Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proceedings of the National Academy of Sciences of the United*

States of America, *86*(6), 1939–1943.

Pääbo, S., Poinar, H., Serre, D., Jaenicke-Despres, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L., & Hofreiter, M. (2004). Genetic analyses from ancient DNA. *Annual Review of Genetics*, *38*, 645–679.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093.

Peter, B. M., Petkova, D., & Novembre, J. (2019). Genetic landscapes reveal how human genetic diversity aligns with geography. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msz280

Peter, B. M., & Slatkin, M. (2013). Detecting range expansions from genetic data. *Evolution; International Journal of Organic Evolution*, *67*(11), 3274–3289.

Petkova, D., Novembre, J., & Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics*, *48*(1), 94–100.

Phillipson, D. W. (2005). *African Archaeology*. Cambridge University Press.

Pickrell, J. K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Güldemann, T., Kure, B., Mpoloka, S. W., Nakagawa, H., Naumann, C., Lipson, M., Loh, P.-R., Lachance, J., Mountain, J., Bustamante, C. D., Berger, B., Tishkoff, S. A., Henn, B. M., Stoneking, M., … Pakendorf, B. (2012a). The genetic prehistory of southern Africa. *Nature Communications*, *3*, 1143.

Pickrell, J. K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Güldemann, T., Kure, B., Mpoloka, S. W., Nakagawa, H., Naumann, C., Lipson, M., Loh, P.-R., Lachance, J., Mountain, J., Bustamante, C. D., Berger, B., Tishkoff, S. A., Henn, B. M., Stoneking, M., … Pakendorf, B. (2012b). The genetic prehistory of southern Africa. *Nature Communications*, *3*, 1143.

Pickrell, J. K., Patterson, N., Loh, P.-R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., & Reich, D. (2014). Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences of the United States of America*,

111(7), 2632–2637.

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, *8*(11), e1002967.

Pinhasi, R., Fernandes, D., Sirak, K., Novak, M., Connell, S., Alpaslan-Roodenberg, S., Gerritsen, F., Moiseyev, V., Gromov, A., Raczky, P., Anders, A., Pietrusewsky, M., Rollefson, G., Jovanovic, M., Trinhhoang, H., Bar-Oz, G., Oxenham, M., Matsumura, H., & Hofreiter, M. (2015). Optimal Ancient DNA Yields from the Inner Ear Part of the Human Petrous Bone. *PloS One*, *10*(6), e0129102.

Plagnol, V., & Wall, J. D. (2006). Possible ancestral structure in human populations. *PLoS Genetics*, *2*(7), e105.

Posth, C., Nakatsuka, N., Lazaridis, I., Skoglund, P., Mallick, S., Lamnidis, T. C., Rohland, N., Nägele, K., Adamski, N., Bertolini, E., Broomandkhoshbacht, N., Cooper, A., Culleton, B. J., Ferraz, T., Ferry, M., Furtwängler, A., Haak, W., Harkins, K., Harper, T. K., … Reich, D. (2018). Reconstructing the Deep Population History of Central and South America. *Cell*, *0*(0). https://doi.org/10.1016/j.cell.2018.10.027

Prendergast, M. E., Lipson, M., Sawchuk, E. A., Olalde, I., Ogola, C. A., Rohland, N., Sirak, K. A., Adamski, N., Bernardos, R., Broomandkhoshbacht, N., Callan, K., Culleton, B. J., Eccles, L., Harper, T. K., Lawson, A. M., Mah, M., Oppenheimer, J., Stewardson, K., Zalzala, F., … Reich, D. (2019). Ancient DNA reveals a multistep spread of the first herders into sub-Saharan Africa. *Science*. https://doi.org/10.1126/science.aaw6275

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., … Pääbo, S. (2014a). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, *505*(7481), 43–49.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., … Pääbo, S.

(2014b). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, *505*(7481), 43–49.

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., & Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(44), 15942–15947.

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L. F., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., … Pääbo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, *468*(7327), 1053–1060.

Savinov, D. G. (2002). *Rannie kochevniki Verkhnego Yeniseya [Early Nomads of Upper Yenisei]*. St. Petersburg State University.

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., & Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PloS One*, *7*(3), e34131.

Schiffels, S., & Durbin, R. (2014a). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*(8), 919–925.

Schiffels, S., & Durbin, R. (2014b). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*(8), 919–925.

Schiffels, S., Haak, W., Paajanen, P., Llamas, B., Popescu, E., Loe, L., Clarke, R., Lyons, A., Mortimer, R., Sayer, D., Tyler-Smith, C., Cooper, A., & Durbin, R. (2016). Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nature Communications*, *7*, 10408.

Schlebusch, C. M., & Jakobsson, M. (2018). Tales of Human Migration, Admixture, and Selection in Africa. *Annual Review of Genomics and Human Genetics*. https://doi.org/10.1146/annurev-genom-083117-021759

Schlebusch, C. M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A. R., Vicente, M., Steyn, M., Soodyall, H., Lombard, M., & Jakobsson, M. (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science*, *358*(6363), 652–655.

Schlebusch, C. M., Skoglund, P., Sjödin, P., Gattepaille, L. M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M. G. B., Soodyall, H., & Jakobsson, M. (2012). Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science*, *338*(6105), 374–379.

Schuenemann, V. J., Peltzer, A., Welte, B., van Pelt, W. P., Molak, M., Wang, C.-C., Furtwängler, A., Urban, C., Reiter, E., Nieselt, K., Teßmann, B., Francken, M., Harvati, K., Haak, W., Schiffels, S., & Krause, J. (2017). Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. *Nature Communications*, *8*, 15694.

Sheehan, S., Harris, K., & Song, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. *Genetics*, *194*(3), 647–662.

Sikora, M., Pitulko, V. V., Sousa, V. C., Allentoft, M. E., Vinner, L., Rasmussen, S., Margaryan, A., de Barros Damgaard, P., de la Fuente, C., Renaud, G., Yang, M. A., Fu, Q., Dupanloup, I., Giampoudakis, K., Nogués-Bravo, D., Rahbek, C., Kroonen, G., Peyrot, M., McColl, H., … Willerslev, E. (2019). The population history of northeastern Siberia since the Pleistocene. *Nature*, *570*, 182–188.

Siska, V., Jones, E. R., Jeon, S., Bhak, Y., Kim, H.-M., Cho, Y. S., Kim, H., Lee, K., Veselovskaya, E., Balueva, T., Gallego-Llorente, M., Hofreiter, M., Bradley, D. G., Eriksson, A., Pinhasi, R., Bhak, J., & Manica, A. (2017). Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago. *Science Advances*, *3*(2), e1601877.

Skoglund, P., Thompson, J. C., Prendergast, M. E., Mittnik, A., Sirak, K., Hajdinjak, M., Salie, T., Rohland, N., Mallick, S., Peltzer, A., Heinze, A., Olalde, I., Ferry, M., Harney,

E., Michel, M., Stewardson, K., Cerezo-Román, J. I., Chiumia, C., Crowther, A., … Reich, D. (2017). Reconstructing Prehistoric African Population Structure. *Cell*, *171*(1), 59–71.e21.

Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genetical Research*, *58*(2), 167–175.

Sousa, V., & Hey, J. (2013). Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews. Genetics*, *14*(6), 404–414.

Steinrücken, M., Kamm, J., Spence, J. P., & Song, Y. S. (2019). Inference of complex population histories using whole-genome sequences from multiple populations. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(34), 17115–17120.

Terhorst, J., Kamm, J. A., & Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, *49*(2), 303–309.

Tishkoff, S. A., Gonder, M. K., Henn, B. M., Mortensen, H., Knight, A., Gignoux, C., Fernandopulle, N., Lema, G., Nyambo, T. B., Ramakrishnan, U., Reed, F. A., & Mountain, J. L. (2007). History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Molecular Biology and Evolution*, *24*(10), 2180–2195.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J.-M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., … Williams, S. M. (2009). The genetic structure and history of Africans and African Americans. *Science*, *324*(5930), 1035–1044.

Tseveendorj, D. (2007). *Chandmanii Soyol [Chandman Culture]*. Mongolian Academy of Sciences.

Vai, S., Brunelli, A., Modi, A., Tassi, F., Vergata, C., Pilli, E., Lari, M., Susca, R. R., Giostra, C., Baricco, L. P., Bedini, E., Koncz, I., Vida, T., Mende, B. G., Winger, D., Loskotová, Z., Veeramah, K., Geary, P., Barbujani, G., … Ghirotto, S. (2019). A genetic perspective

on Longobard-Era migrations. *European Journal of Human Genetics: EJHG*, *27*(4), 647–656.

van de Loosdrecht, M., Bouzouggar, A., Humphrey, L., Posth, C., Barton, N., Aximu-Petri, A., Nickel, B., Nagel, S., Talbi, E. H., El Hajraoui, M. A., Amzazi, S., Hublin, J.-J., Pääbo, S., Schiffels, S., Meyer, M., Haak, W., Jeong, C., & Krause, J. (2018). Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations. *Science*, *360*(6388), 548–552.

Veeramah, K. R., Rott, A., Groß, M., van Dorp, L., López, S., Kirsanow, K., Sell, C., Blöcher, J., Wegmann, D., Link, V., Hofmanová, Z., Peters, J., Trautmann, B., Gairhos, A., Haberstroh, J., Päffgen, B., Hellenthal, G., Haas-Gebhard, B., Harbeck, M., & Burger, J. (2018). Population genomic analysis of elongated skulls reveals extensive female-biased immigration in Early Medieval Bavaria. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(13), 3494–3499.

Wang, C.-C., Reinhold, S., Kalmykov, A., Wissgott, A., Brandt, G., Jeong, C., Cheronet, O., Ferry, M., Harney, E., Keating, D., Mallick, S., Rohland, N., Stewardson, K., Kantorovich, A. R., Maslov, V. E., Petrenko, V. G., Erlikh, V. R., Atabiev, B. C., Magomedov, R. G., … Haak, W. (2019). Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. *Nature Communications*, *10*(1), 590.

Wang, K., Mathieson, I., O'Connell, J., & Schiffels, S. (2020). Tracking human population structure through time from whole genome sequences. *PLoS Genetics*, *16*(3), e1008552.

Wang, S., Lachance, J., Tishkoff, S. A., Hey, J., & Xing, J. (2013). Apparent variation in Neanderthal admixture among African populations is consistent with gene flow from Non-African populations. *Genome Biology and Evolution*, *5*(11), 2075–2081.

Wang, Y., & Hey, J. (2010). Estimating divergence parameters with small samples from a large number of loci. *Genetics*, *184*(2), 363–379.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen,

Y.-J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X.-Z., Liu, Y., … Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, *452*(7189), 872–876.

Wiuf, C., & Hein, J. (1999). Recombination as a point process along sequences. *Theoretical Population Biology*, *55*(3), 248–259.

Xu, D., Pavlidis, P., Taskent, R. O., Alachiotis, N., Flanagan, C., DeGiorgio, M., Blekhman, R., Ruhl, S., & Gokcumen, O. (2017). Archaic Hominin Introgression in Africa Contributes to Functional Salivary MUC7 Genetic Variation. *Molecular Biology and Evolution*, *34*(10), 2704–2715.

Zischler, H., Höss, M., Handt, O., von Haeseler, A., van der Kuyl, A. C., & Goudsmit, J. (1995). Detecting dinosaur DNA [Review of *Detecting dinosaur DNA*]. *Science*, *268*(5214), 1192–1193; author reply 1194.

9. Summary

The revolution of sequencing technology has brought an exponential increase in the production of genomic data. This thesis tackles global and continental questions on human demographic history from two directions using genetic data. Manuscript A provides a novel analytical method for estimating migration rate and effective population size utilizing high-coverage whole genome sequences, while manuscript B and C reveal the history of population movement and interactions by directly analyzing genome-wide data from ancient individuals.

First, I developed a new method called MSMC-IM for deciphering population interactions quantitatively using whole genome sequence data from modern humans (**Manuscript A**). This new method estimates migration rates and effective population sizes over time for a pair of populations. Based on within-/across-population coalescence rates calculated from genomic data using the so-called Multiple Sequentially Markovian Coalescent method (MSMC), the new method MSMC-IM fits a continuous Isolation-Migration model to the coalescence rates, which assumes two separate populations connected via a time-dependent migration rate. I implemented this approach and tested it with simulated data from different demographic scenarios involving post-split admixture or archaic introgression. Applying the method to the genomes from 15 worldwide human populations, MSMC-IM reveals the process of worldwide pairwise separation from a few thousand up to several million years ago. In particular, MSMC-IM detects extremely deep ancestry in present-day African populations, such as the southern African San and the central African Mbuti, with a proportion in their genome tracing back to a million years ago and beyond.

Second, I analyzed genome-wide data from 20 newly reported ancient individuals in sub-Saharan Africa (**Manuscript B**). These individuals are dated to between 4000 to 200 years ago, and associated with all key subsistence strategies in Africa - foraging, herding and farming. Combined with published African aDNA, the new results suggest a contraction of hunter-gatherer ancestry in the past in eastern Africa which used to be widespread in central, eastern and southern Africa, and also suggests that coexistence and interactions between foragers and herders in eastern Africa in the last four thousand years was more complex than previously reported. The new results record the arrival of Nilotic-related and Bantu-related ancestry in eastern Africa during the Iron Age and show the expansion of eastern African pastoralist-related ancestry to central Africa during the same time period. Newly reported ancient individuals also directly document the earlier arrival of eastern pastoralists-related ancestry than Banu-related ancestry in Botswana at around the first millennium, and the presence of Bantu-related ancestry in the western coastline of the Democratic Republic of the Congo hundred years ago.

Third, I analyzed genome-wide data from 214 newly reported ancient individuals in the Eastern Steppe (**Manuscript C**). This dataset records the dynamic changes of the genetic profile in Mongolia and surrounding regions in Russian from the pre-Bronze Age (ca. 4600 BCE) to the Mongol empire (ca. 1400 CE). This time transect over 6000 years covers major demographic events associated with subsistence changes and formation of well-known pastoral empires centered on Mongolia. Before the introduction of pastoralism, ancestry represented by hunter-gatherers from the Devil's Cave in the Far East, was widely prevalent in a large territory ranging from Lake Baikal, through Mongolia, to the Russian Far East.

From the Early Bronze Age, western herders, such as the Afanasievo, started to emerge in central Mongolia and expanded up to 1500km further east. However, the subsequent Chemurchek culture in the Altai mountain did not show the genetic link to Afansievo that was previously hypothesized from cultural similarity, but represents new migration waves. During the Middle and Late Bronze Age and Iron Age, the ancient people analysed here show strong spatial-genetic correlation in western, northern and eastern Mongolia, who derive their ancestry from a new MLBA central Steppe population - Sintashta - to various degrees. Later, the formation of the Xiongnu, the first empire of nomadic pastoralists, was associated with the mixture of these previously separated populations in western and eastern Mongolia, and several rapid new gene influxes from various regions across the Eurasia continent. The genetic heterogeneity continues in the subsequent Turkic and Uigur empires of the early Medieval period. From the Mongol empire period, sampled individuals show a remarkable increase in eastern Eurasian ancestry, marking the first appearance of the genetic profile that we see in Mongolia today.

All in all, this thesis demonstrates how modern and ancient genomic data helps reconstructing the history of human population separations, movements and admixture. For the future work, one direction is to develop more new computational methods to extract more information from available genomic data - either modern or ancient. Another would be to sample more ancient individuals from a wider spatial and temporal range to fill the current sampling gaps and broaden the overall picture of human history. Studies in both directions would enlarge our sphere of understanding population genetics.

10. Zusammenfassung

Entwicklungen der vergangenen 10 Jahre im Bereich *DNA Sequencing* haben zu einem exponentiellen Anstieg neuer genetischer Daten geführt. Das gilt sowohl für komplette Genomsequenzen (über *Shotgun*-Sequenzierung) als auch für genomweite Daten auf herkunfts-informativen SNPs (über *in-solution hybridization capture*). Die vorliegende Arbeit behandelt Fragen zur menschlichen Populationsgeschichte auf globalem und kontinentalem Maßstab. Dabei kommen beide genannten Arten von Daten zum Einsatz. Manuskript A führt eine neuartige analytische Methode zur Schätzung von Migrationsraten und der effektiven Populationsgröße unter Verwendung von hochauflösenden Gesamtgenomsequenzen ein, während die Manuskripte B und C konkrete Ereignisse von Bevölkerungsbewegung und -interaktion anhand von *low-coverage SNP capture Daten* von prähistorischen Individuen rekonstruieren.

Für **Manuskript A** habe ich eine neue, populationsgenetische Methode entwickelt – MSMC-IM. Sie kann dafür genutzt werden, Fragen zur menschlichen Demographieentwicklung unter Verwendung vollständiger, moderner Genomsequenzen zu beantworten. MSMC-IM liefert Schätzungen zur Migrationsrate und zur diachronen Entwicklung der effektiven Bevölkerungsgröße für Populationspaare. Basierend auf den mittels MSMC berechneten Koaleszenzraten innerhalb und zwischen Bevölkerungen passt die neue Methode MSMC-IM ein kontinuierliches Isolation-Migrationsmodell an die Koaleszenzraten an, das von zwei sich allmählich trennenden Populationen mit durch die Zeit variabler Migrationsrate ausgeht. Ich habe diesen Ansatz implementiert und ihn mit simulierten Daten verschiedener, komplexer demographischer Szenarien getestet. Die Testszenarien umfassten unter anderem verschiedene Formen von *post-split* Vermischung (z.B. *archaic introgression*). Bei der Anwendung auf 15 menschliche Populationen aus verschiedenen Regionen der Welt rekonstruiert MSMC-IM einen Prozess weltweiter, paarweiser Trennung über einen Zeitraum von einigen tausend bis zu mehreren Millionen Jahren. Insbesondere für einige rezente afrikanische Populationen weist MSMC-IM extrem tiefe Ancestry nach: Sowohl für die südafrikanischen San als auch die zentralafrikanischen Mbuti reicht der Prozess der Populationsdivergenz bis zu eine Million Jahre und länger zurück.

Manuskript B verarbeitet genomweite Daten von 20 neu beprobten, alten Individuen aus Subsahara-Afrika. Diese Individuen datieren in ein Zeitfenster von vor fast viertausend bis zweihundert Jahren vor heute und decken alle für den afrikanischen Kontinent wichtigen Subsistenzstrategien – Jagd, Sammeln, Viehzucht und Ackerbau – ab. Zusammen mit bereits veröffentlichten afrikanischen aDNA-Daten zeigen die neuen Ergebnisse einen zeitweisen Rückgang jener genetischen Herkunft alter Jäger-Sammler-Gruppen in Ostafrika, die heute in Zentral-, Ost- und Südafrika wieder weit verbreitet ist. Sie deuten auch darauf hin, dass Zusammenleben und Interaktion von Jägern und Hirten in Ostafrika in den letzten viertausend Jahren komplexer ablief als bisher angenommen. Die neuen Ergebnisse erfassen auch eine neue Abstammungskomponente mit Bezügen zum nilotischen- und dem Bantu-Raum im eisenzeitlichen Ostafrika. Außerdem legen sie nahe, dass die Ausbreitung ostafrikanischer Hirten-Herkunft im gleichen Zeitraum bis nach Zentralafrika reicht. Neu beprobte Individuen belegen die frühere Ankunft der östlichen Hirten Ancestry in Botswana im ersten Jahrtausend gegenüber der späteren Bantu Ancestry. Schließlich konnte Bantu Ancestry an der Westküste der DR Kongo vor nur einhundert Jahren nachgewiesen werden.

Für **Manuskript C** habe ich genomweite Daten von 214 neu beprobten, bis zu 6000 Jahre alten Individuen aus dem östlichen Steppenraum analysiert. Dieser Datensatz erfasst die dynamischen Veränderungen des genetischen Profils im Areal der heutigen Mongolei und den umliegenden Regionen im heutigen Russland ausgehend von einem Zeithorizont vor Beginn der Bronzezeit (ca. 4600 v.Chr.) bis zum Mongolischen Reich (ca. 1400 n.Chr.). Dieser zeitliche Transekt deckt wichtige demographische Ereignisse ab: Veränderungen in der Subsistenzstrategie und die Entstehung der bekannten mongolischen Reiternomaden Reiche. Vor der Einführung der Viehzucht war neolithische Jäger- und Sammler-Abstammung – vertreten durch Individuen aus dem Fundort Devils Cave im äußersten Osten des Untersuchungsgebiets – dominant. Das Verbreitungsgebiet dieser Abstammungskomponente reichte vom Baikalsee über die Mongolei bis zum fernen Osten Russlands. Ab der frühen Bronzezeit tauchten westlichen Hirtengruppen, wie z.B. Afanasievo, in der zentralen Mongolei auf und breiteten sich bis zu 1500km weiter nach Osten aus (Afanasievo-Expansion). Die darauf folgende Chemurchek-Kultur im Altai kann jedoch nicht, wie zunächst aufgrund kultureller Ähnlichkeiten angenommen, direkt mit Afansievo Ancestry in Verbindung gebracht werden, sondern verkörpert neue Migrationswellen. Für die mittlere und späte Bronze- und Eisenzeit zeigen die für diese Studie beprobten Individuen eine starke räumlich-genetische Korrelation in der westlichen, nördlichen und östlichen Mongolei. Lokale Gruppen sind klar voneinander separierbar. Sie tragen in unterschiedlichem Umfang Abstammung der Sintashta, einer mittelbronzezeitlichen Zentralsteppenpopulation. Die Entstehung Xiongnus, des ersten mongolischen Reiternomadeneiches, geht mit der Vermischung dieser zuvor getrennten Populationen in der westlichen und östlichen Mongolei und einem raschen neuen Gen-Zustrom aus dem gesamten eurasischen Kontinent einher. Die genetische Heterogenität setzt sich in den nachfolgenden Turkischen und Uigurischen Reichen des Frühmittelalters fort. In der Zeit des Mongolischen Reiches zeigen die untersuchten Individuen eine bemerkenswerte Zunahme osteurasischer Ancestry. Das markiert das erste Auftreten eines genetischen Profils, das dem heutiger Mongolen ähnelt.

Alles in allem zeigt die vorliegende Arbeit, wie moderne und alte Genomdaten helfen können, die Geschichte von Populationstrennungen, -bewegungen und -vermischungen zu rekonstruieren. Eine Perspektive ist die Entwicklung neuer Methoden um aus den verfügbaren Genomdaten – sowohl moderne als auch alte – Wissen abzuleiten. Angesichts des sich erweiternden aDNA-Datenbestandes sind neue Methoden erforderlich. Daneben ist es erforderlich, mehr alte Individuen aus einem größeren räumlichen und zeitlichen Spektrum zu beproben, um Verzerrungen durch Verteilungsschwerpunkte (sampling bias) zu verringern und die Lücken im Gesamtbild der menschlichen Geschichte zu füllen. Zukünftige Foschung muss beide Perspektiven beachten um unser Verständnis menschlicher Populationsgenetik zu vertiefen.

11. Eigenständigkeitserklärung

Entsprechend §5 Abs. 4 der Promotionsordnung der Biologisch-Pharmazeutischen Fakultät der Friedrich-Schiller-Universität Jena, erkläre ich, dass mir die geltende Promotionsordnung der Fakultät bekannt ist. Ich bezeuge, dass ich die vorliegende Dissertation selbst angefertigt habe und keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönliche Mitteilungen sowie Quellen in meiner vorliegenden Arbeit angegeben habe. Zudem habe ich alle Personen, die mir bei der Auswahl und Auswertung sowie bei der Erstellung der Manuskripte unterstützt haben, in der Auflistung der Manuskripte und den entsprechenden Danksagungen namentlich erwähnt. Zudem verichere ich, dass ich die Hilfe eines Promotionsberaters nicht in Anspruch genommen haben und auch Dritten von mir keine unmittelbaren sowie mittelbaren geldwerte Leistungen für Arbeiten, die im Zusammenhang mit dieser Dissertaton stehen, erhalten haben. Die vorliegende Promotion wurde zuvor weder für eine staatliche oder andere wissenschaftliche Prüfung eingereicht, also auch einer anderen Hochschule als Dissertation vorgelegt.

Jena, den 30.3.2020

_____

Ke Wang

12. Acknowledgments

I would like to thank the following people for their help and support along this fantastic doctorate journey.

To my supervisor Stephan Schiffels, for your thorough guidance in doing research, for teaching me to be an open-minded scientist, for giving me the freedom to grow independently, for being a role model not just in terms of as a scientist but also as a responsible parent.

To Choongwon Jeong, for providing me the chance of working on Mongolia project, for mentoring me through the entire project and all your kind advice on my career plan. To Tina Warinner, for your trust in the project. It is my greatest luck to work closely with you two.

To Johannes Krause for the opportunity to work in such an outstanding research institute. To Wolfgang Haak, Cosimo Posth and all participants in our weekly popgen meeting, for the warm atmosphere and all fruitful discussions we had.

To my archeologist colleagues and friends, for sharing enormous archeology knowledge and showing me a new angle to view my own research.

To my dear friends in Jena, Eirini, Vannesa, Aida, James, Thiseas, Marcel, for your open ear and support along this journey. You make my life in Jena brighter. Special thanks to Clemens, my dear officemate who helped me with writing Zusammenfassung, and to James, who helped me on checking the grammar of this thesis and offered me many helpful comments.

To my little brother, for all the joy you have brought to my life and making me a responsible person. To my father, for teaching me to be kind and full of justice, and also for trying hard to discuss the latest research findings with me. To my mama, for your support on pursuing a PhD degree abroad and your accompany over phone day by day, for your understanding on every decision I made so far. Home is behind, but you are in the softest part of my heart.

13. Curriculum Vitae

Personal Information	
Name	Ke Wang
Date of Birth	22.11.1996
School Education	
2020-present	Postdoc in the Department of Archaeogenetics, MPI-SHH, Jena, Germany
2016-2020	PhD in *Genetics*, The Max Planck Institute for the Science of Human History (MPI-SHH) Thesis titled "**Investigating human population history with new computational methods and ancient DNA data"** supervised by Dr. Stephan Schiffels in Archaeogenetics Department, MPI-SHH, Jena, Germany
2015-2016	Msc in *Genetics of human diseases*, University College London (UCL) Thesis titled "**Inferring genetic regions undergoing selection among world-wide human populations**" supervised by Dr. Garrett Hellenthal, Department of Genetics, Evolution & Environment, UCL, London, UK
2011-2015	Bsc in *Biotechnology*, Shandong University, Jinan, China
2008-2011	High school diploma from Yantai No.2 Middle School, Shandong, China
Publications	
Wang, K., Mathieson, I., O'Connell, J., Schiffels, S. (2020) Tracking human population structure through time from whole genome sequences. *PLoS Genet* 16(3): e1008552. https://doi.org/10.1371/journal.pgen.1008552	
Schiffels, S., **Wang, K**. (2020) MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent. In: Dutheil J. (eds) Statistical Population Genomics. *Methods in Molecular Biology*, vol 2090. Humana, New York, NY	
Wang, K.*, Goldstein, S.*, Bleasdale, M., Clist, B., Bostoen, K., Bakwa-Lufu, P., Buck, L. T., Crowther, A., Dème, A., McIntosh, R. J., Mercader, J., Ogola, C., Power, R. C., Sawchuk, E., Robertshaw, P., Wilmsen, E. N., Petraglia, M., Ndiema, E., Manthi, F. K., Krause, J., Roberts, P., Boivin, N., Schiffels, S. (2020). Ancient genomes reveal complex patterns of population movement, interaction and replacement in sub-Saharan Africa. *Science Advances,* 2020;6: eaaz0183. *equal contribution	
Jeong, C.*, **Wang, K.***, Wilkin, S., Taylor, W. T. T., Miller, B. K., Ulziibayar, S., Stahl, R., Chiovelli, C., Bemmann, J. H., Knolle, F., Kradin, N., Bazarov, B. A., Miyagashev, D. A., Konovalov, P. B., Zhambaltarova, E., Miller, A. V., Haak, W., Schiffels, S., Krause, J., Boivin, N., Myagmar, E., Hendy, J., Warinner, C. (2020) A dynamic 6,000-year genetic history of Eurasia's Eastern Steppe. *Cell,* 2020;183: 890–904.e29. *equal contribution	
Ning, C., Li, T., **Wang, K.**, Zhang, F., Li, T., Wu, X., Gao, S., Zhang, Q., Zhang, H., Hudson, M. J., Dong, G., Wu, S., Fang, Y., Liu, C., Feng, C., Li, W., Han, T., Li, R., Wei, J.,	

Zhu, Y., Zhou, Y., Wang, C., Fan, S., Xiong, Z., Sun, Z., Ye, M., Sun, L., Wu, X., Liang, F., Cao, Y., Wei, X., Zhu, H., Zhou, H., Krause, J., Robbeets M., Jeong C., Cui, Y. (2020) Genomic history of northern China for the last 7,500 years. *Nature Communications,* **11,** 2700 (2020).
Teaching and Conferences
Talk titled "*Investigating East African population structure through ancient genomes*" at Conference, International conference - "Africa, the cradle of human diversity", May 2019, Uppsala, Sweden
Talk titled "*Reconstructing population separation history from whole genome sequences*" at EMBO | EMBL Symposium: Reconstructing the Human Past - Using Ancient and Modern Genomics, March 2019, Heidelberg, Germany
Teaching "Coalescent Theory" in DAG Internal BootCamp, MPI-SHH, Jena, Germany
Poster titled "*Fitting an isolation-migration model to MSMC estimates to infer population sizes and migration rates over time*" at SMBE 2018, Yokohama, Japan
Poster titled "*Inferring population split times and rates of gene flow between populations from multiple genome sequences*" at SMBE 2017, Austin, USA

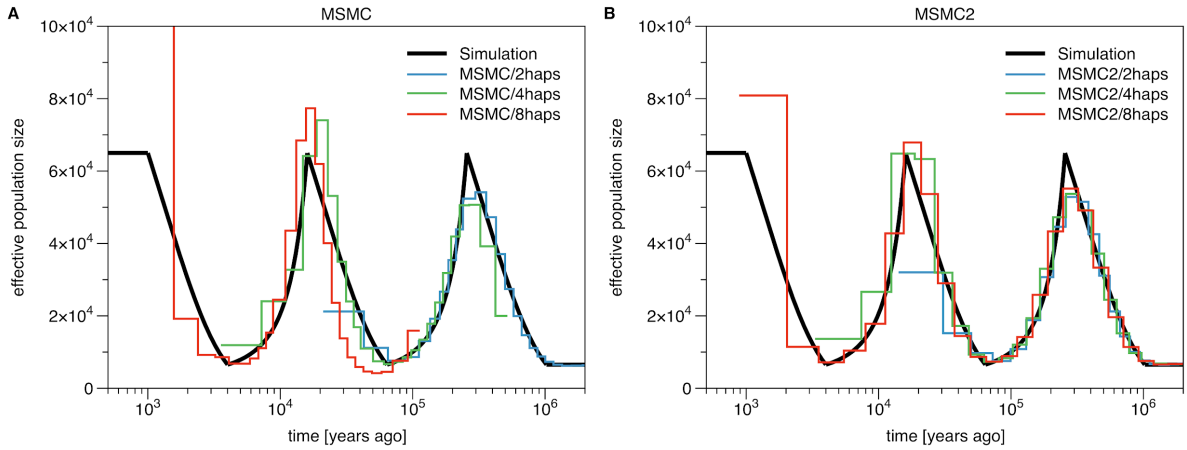
_____

Ke Wang

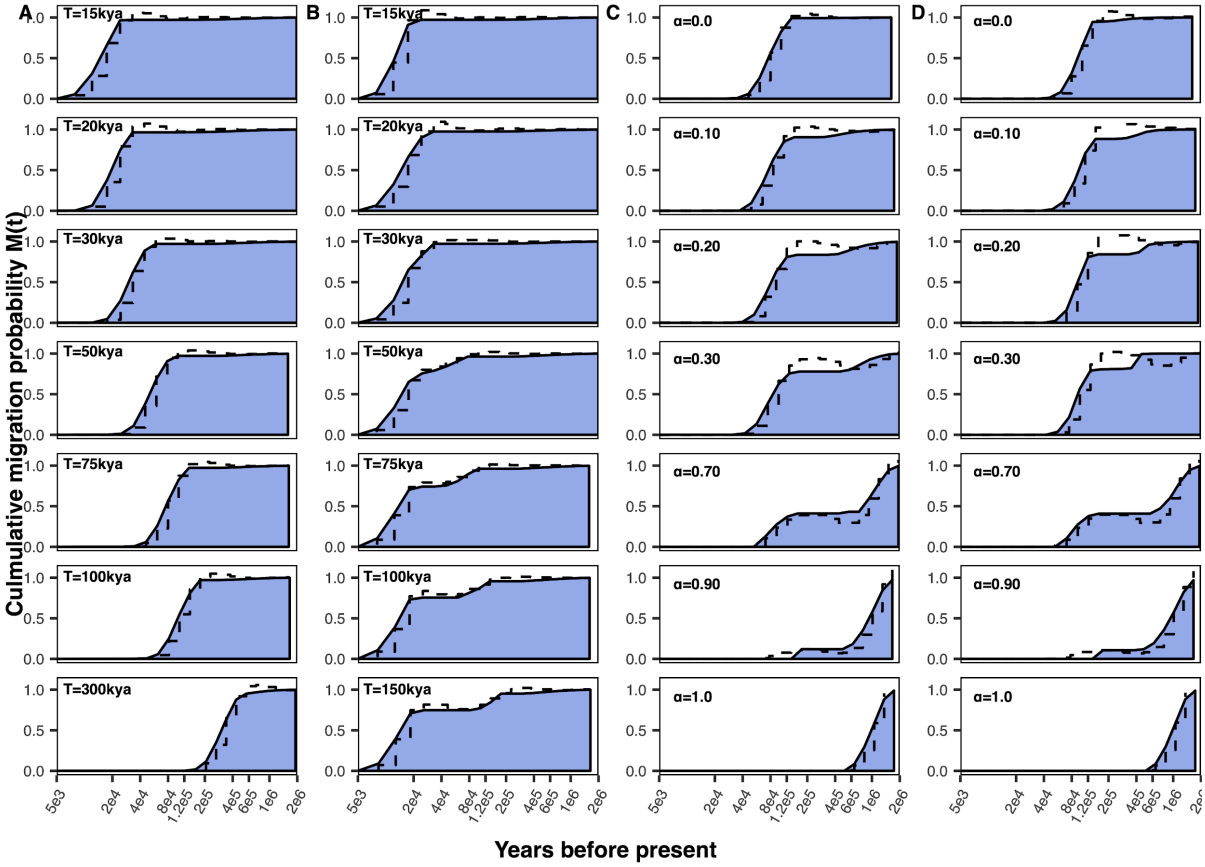
14. Appendix

14.1. Supplementary Materials of paper A

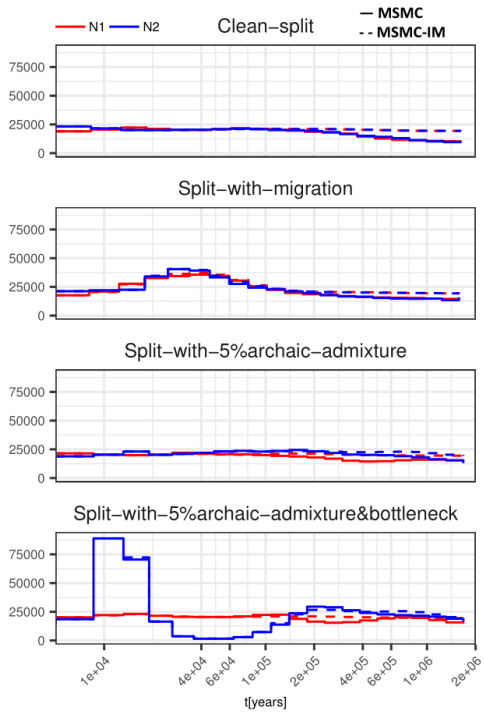
Supplementary Figures and Tables



S1 Fig. MSMC and MSMC2 population size estimates from simulated data.
To test population size inference capabilities of MSMC (A) and MSMC2 (B) applied to two, four and eight haplotypes, we simulated a series of exponential population growths and declines, each changing the population size by a factor ten. The true population size is shown as dark solid line. Compared to MSMC, MSMC2 recovers the population size well, and the resolution in recent times increases with the number of haplotypes. With two haplotypes, MSMC2 infers the population history from 10kya to 3 million years, whereas, with four haplotypes and eight haplotypes the resolution in recent times is extended to 3kya and 1kya years ago respectively.



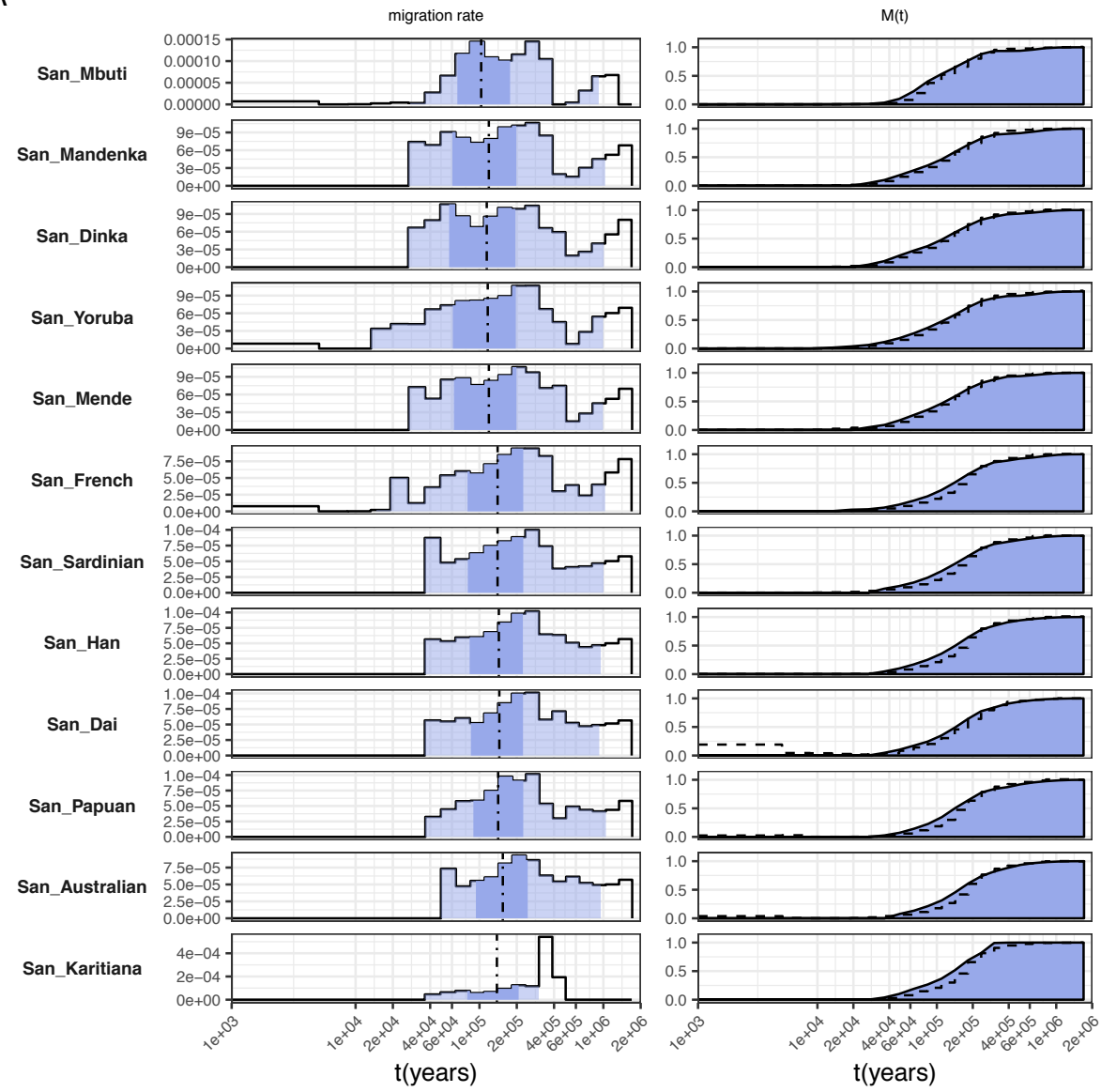
S2 Fig. Cumulative migration probabilities from four simulation scenarios.
This figure shows the same results as Fig 2, but showing M(t) instead of m(t). The scenarios are (A) the *Clean-split* scenario. (B) the *Split-with-migration* scenario, and (C) the *Split-with-archaic-admixture* scenario. (D) the *Split-with-archaic-admixture-and-bottleneck* scenario. For panel (C) and (D), we show results with alpha ranging from 0 to 1, instead of between 0 to 20% shown in Figure 2. The relative CCR is shown in step-wise dashed lines to be compared with M(t).



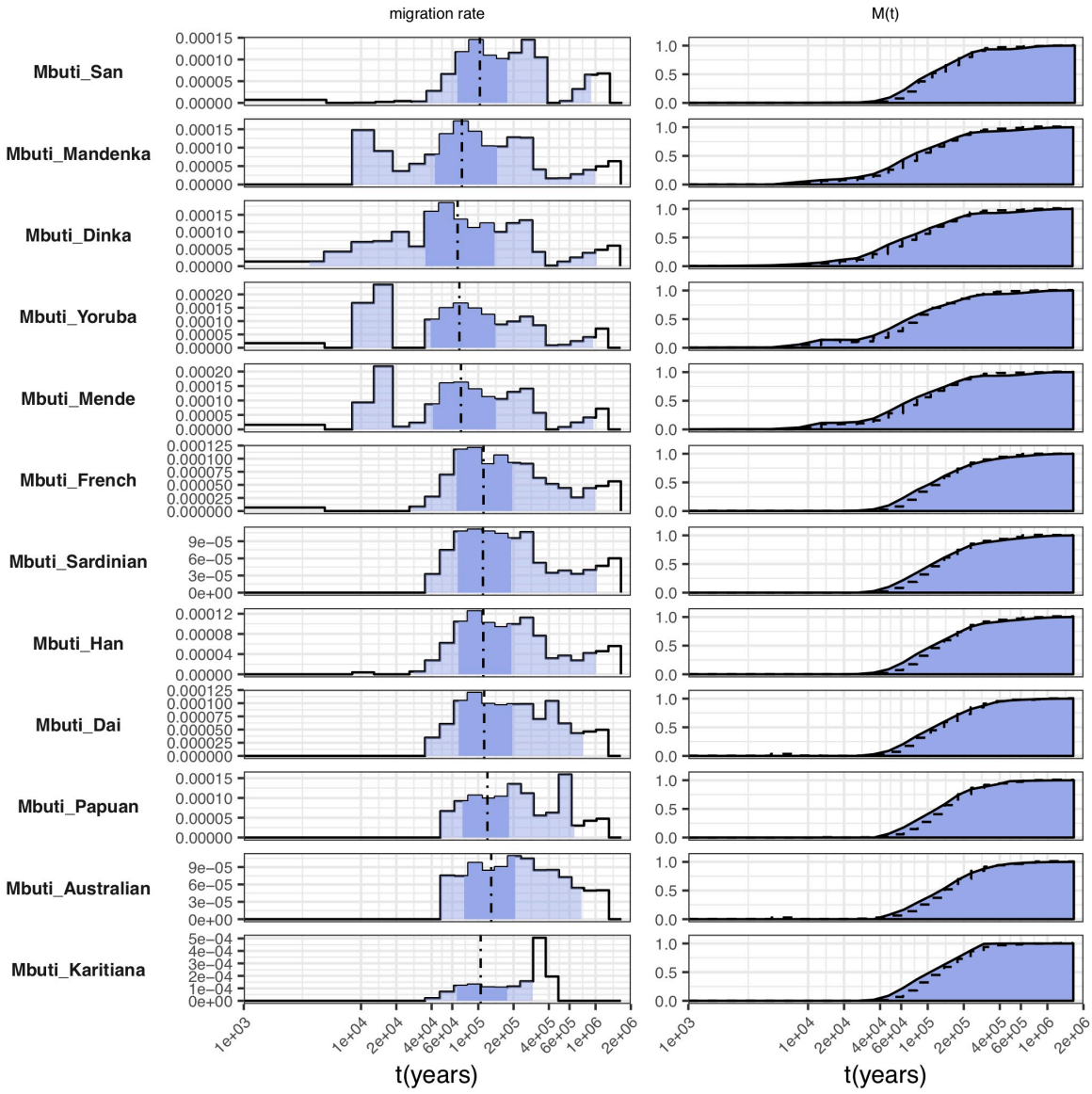
S3 Fig. Population size estimates from MSMC2 compared to MSMC-IM: **we simulated $N_1(t)$ and $N_2(t)$ as constant 20,000 in top three different simulation scenarios, and simulated a severe bottleneck in $N_2(t)$ with a factor 30 between 40-60kya in the bottom simulation scenario.**

The split time *T* is 75kya in all four cases, and all other parameters are the same as in Figure 2 and as indicated. As shown, the MSMC-IM estimates for $N_1(t)$ and $N_2(t)$ are close to the inverse coalescence rates, with relatively small effects caused by the migration rate in MSMC-IM which is absent from MSMC2.

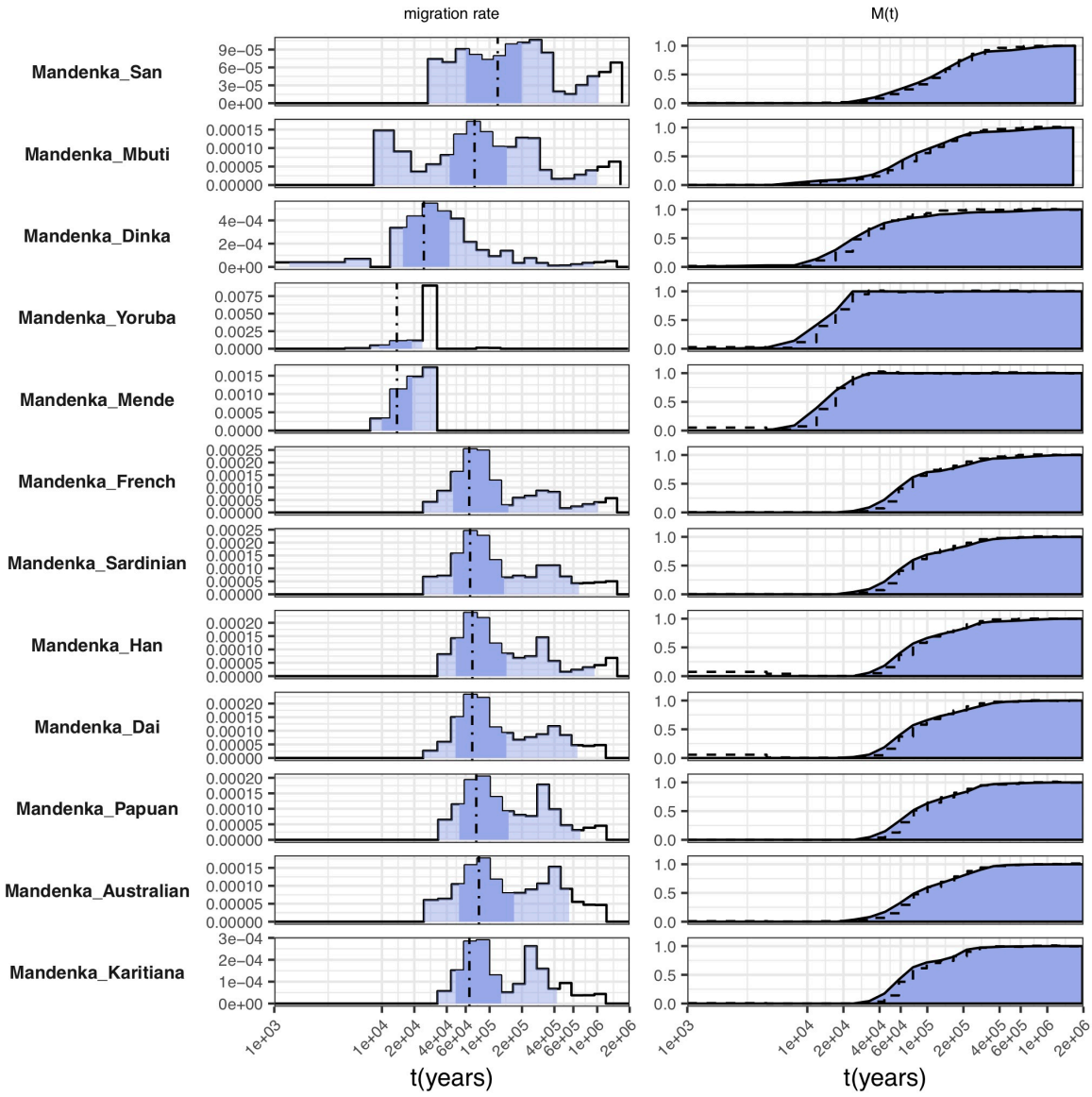
A


migration rate
M(t)
San_Mbuti
San_Mandenka
San_Dinka
San_Yoruba
San_Mende
San_French
San_Sardinian
San_Han
San_Dai
San_Papuan
San_Australian
San_Karitiana
0.00015
0.00010
0.00005
0.00000
9e−05
6e−05
3e−05
0e+00
7.5e−05
5.0e−05
2.5e−05
0.0e+00
1.0e−04
4e−04
2e−04
1.0
0.5
0.0
1e+03
1e+04
2e+04
4e+04
6e+04
1e+05
2e+05
4e+05
6e+05
1e+06
2e+06
t(years)
t(years)

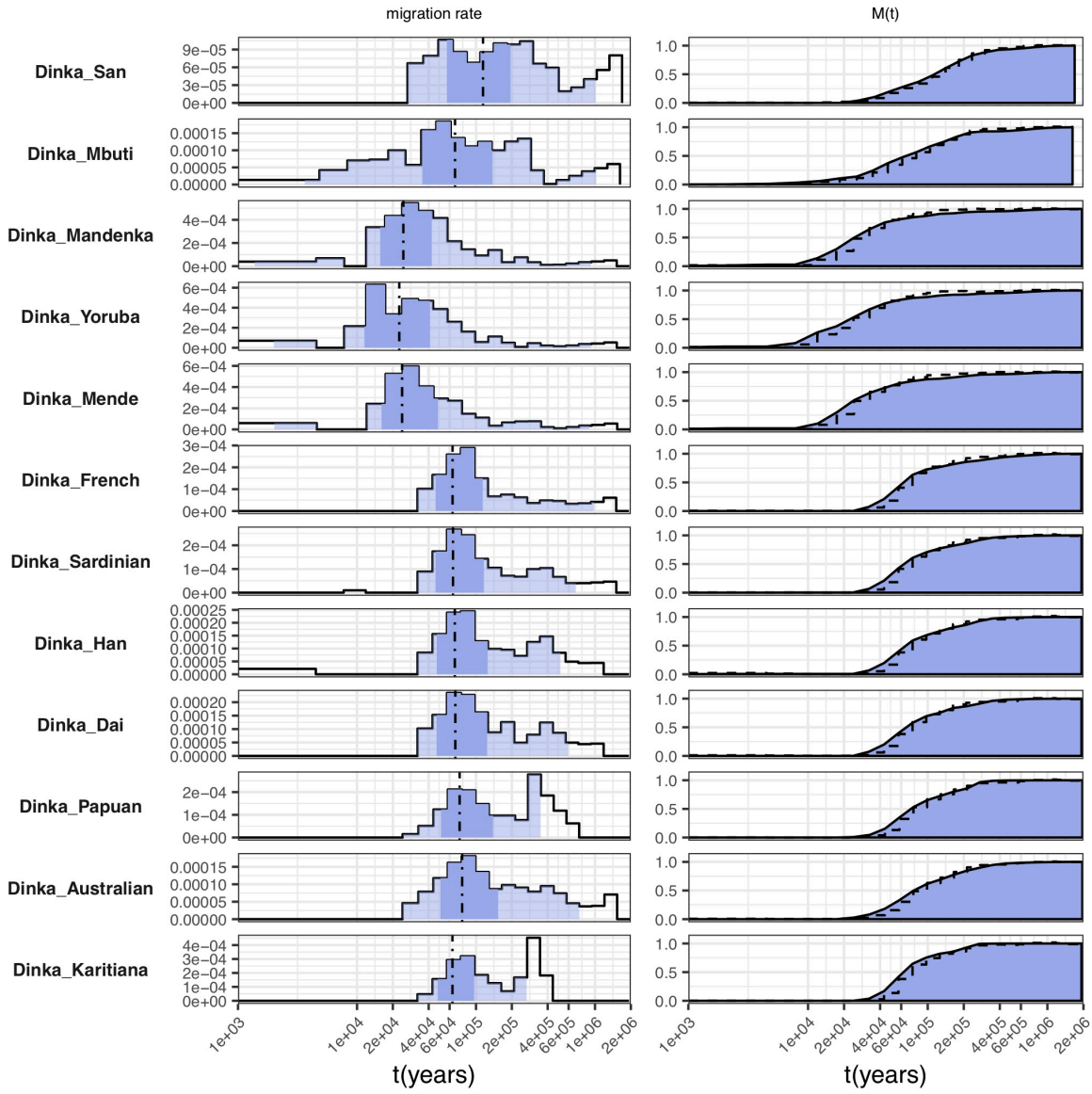
B


migration rate
M(t)
Mbuti_San
Mbuti_Mandenka
Mbuti_Dinka
Mbuti_Yoruba
Mbuti_Mende
Mbuti_French
Mbuti_Sardinian
Mbuti_Han
Mbuti_Dai
Mbuti_Papuan
Mbuti_Australian
Mbuti_Karitiana
t(years)
t(years)

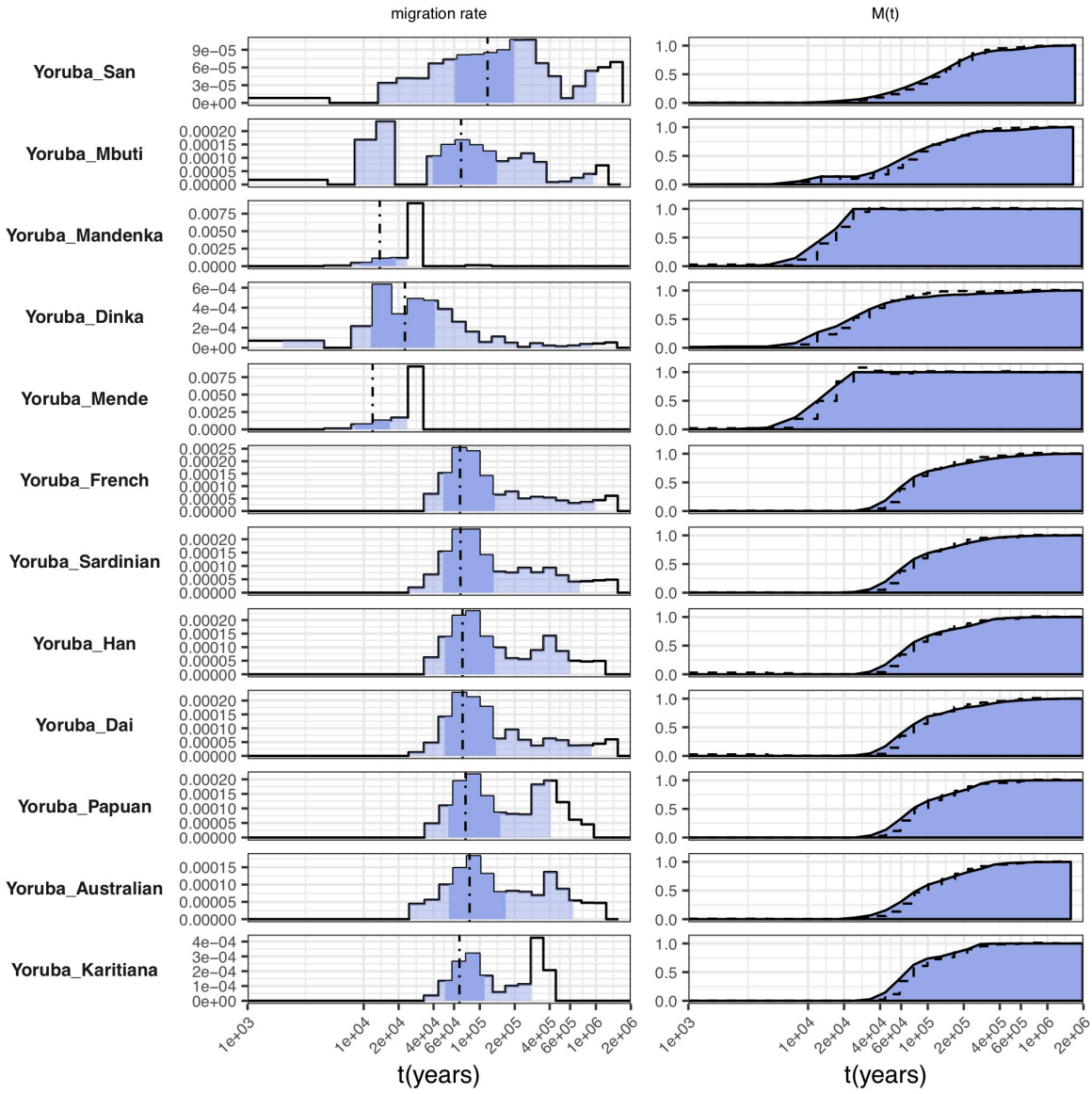
C


migration rate
M(t)
Mandenka_San
Mandenka_Mbuti
Mandenka_Dinka
Mandenka_Yoruba
Mandenka_Mende
Mandenka_French
Mandenka_Sardinian
Mandenka_Han
Mandenka_Dai
Mandenka_Papuan
Mandenka_Australian
Mandenka_Karitiana
t(years)
t(years)

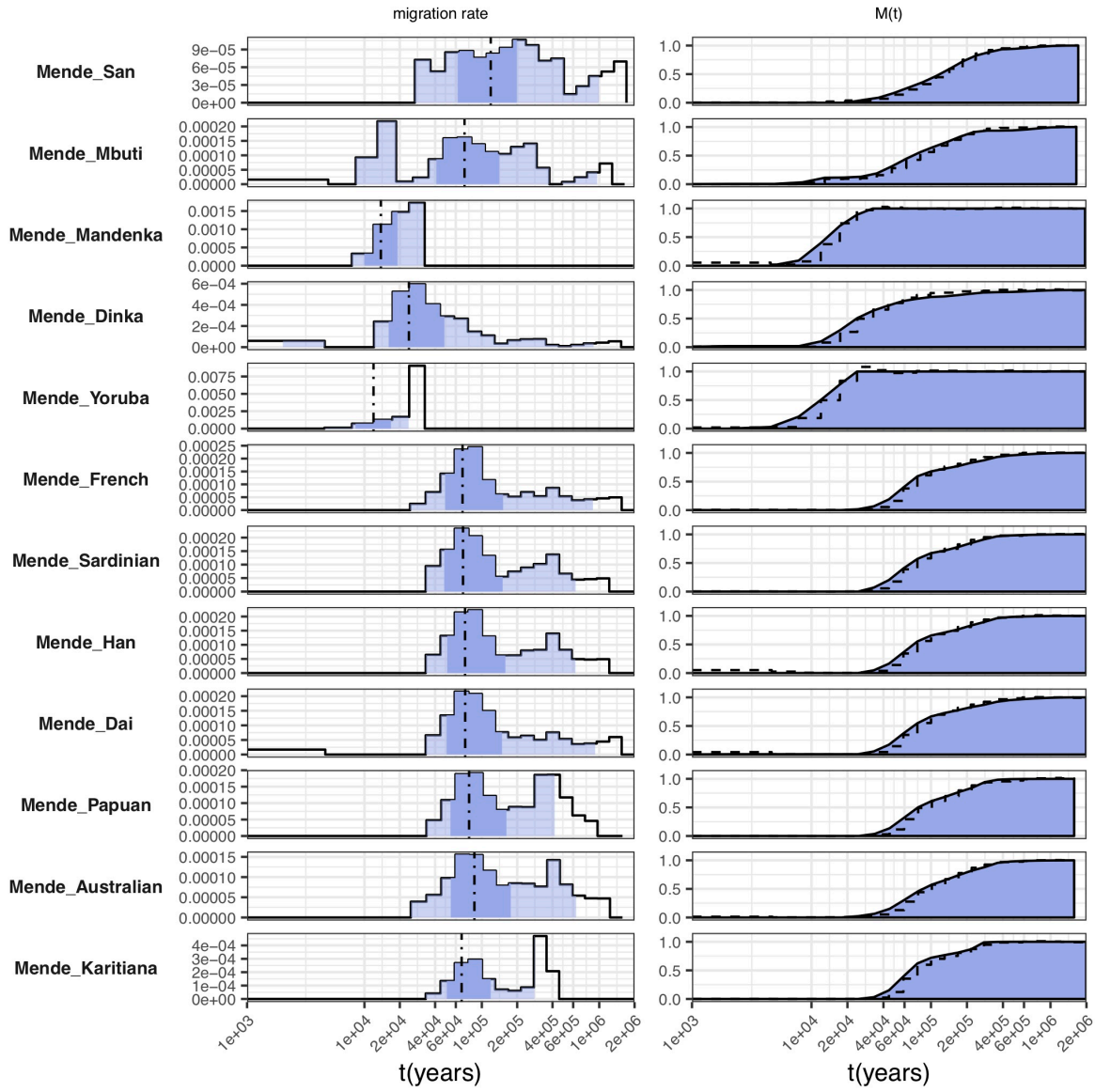
D


migration rate
M(t)
Dinka_San
Dinka_Mbuti
Dinka_Mandenka
Dinka_Yoruba
Dinka_Mende
Dinka_French
Dinka_Sardinian
Dinka_Han
Dinka_Dai
Dinka_Papuan
Dinka_Australian
Dinka_Karitiana
t(years)
t(years)

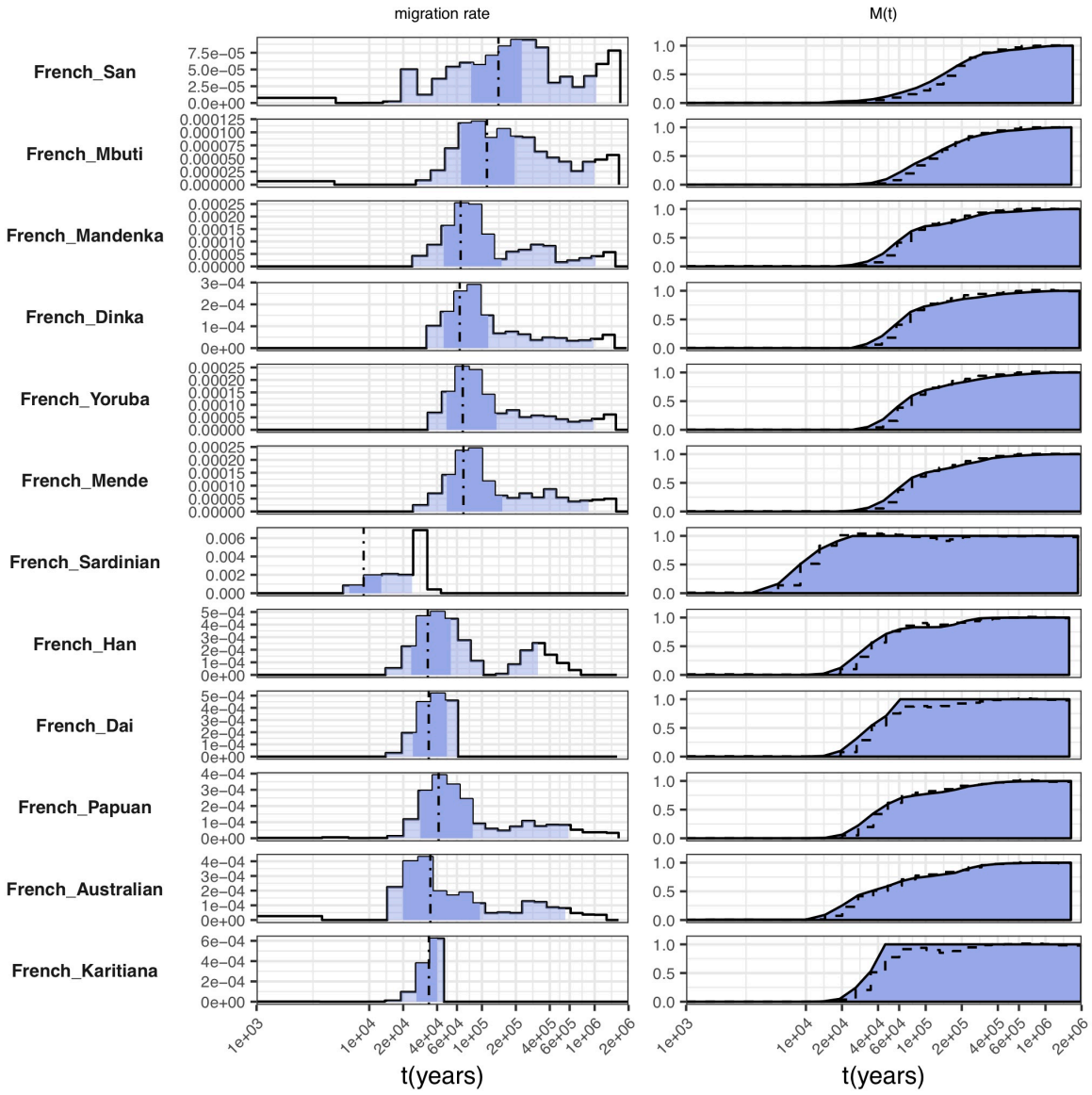
E


migration rate
M(t)
Yoruba_San
Yoruba_Mbuti
Yoruba_Mandenka
Yoruba_Dinka
Yoruba_Mende
Yoruba_French
Yoruba_Sardinian
Yoruba_Han
Yoruba_Dai
Yoruba_Papuan
Yoruba_Australian
Yoruba_Karitiana
t(years)
t(years)

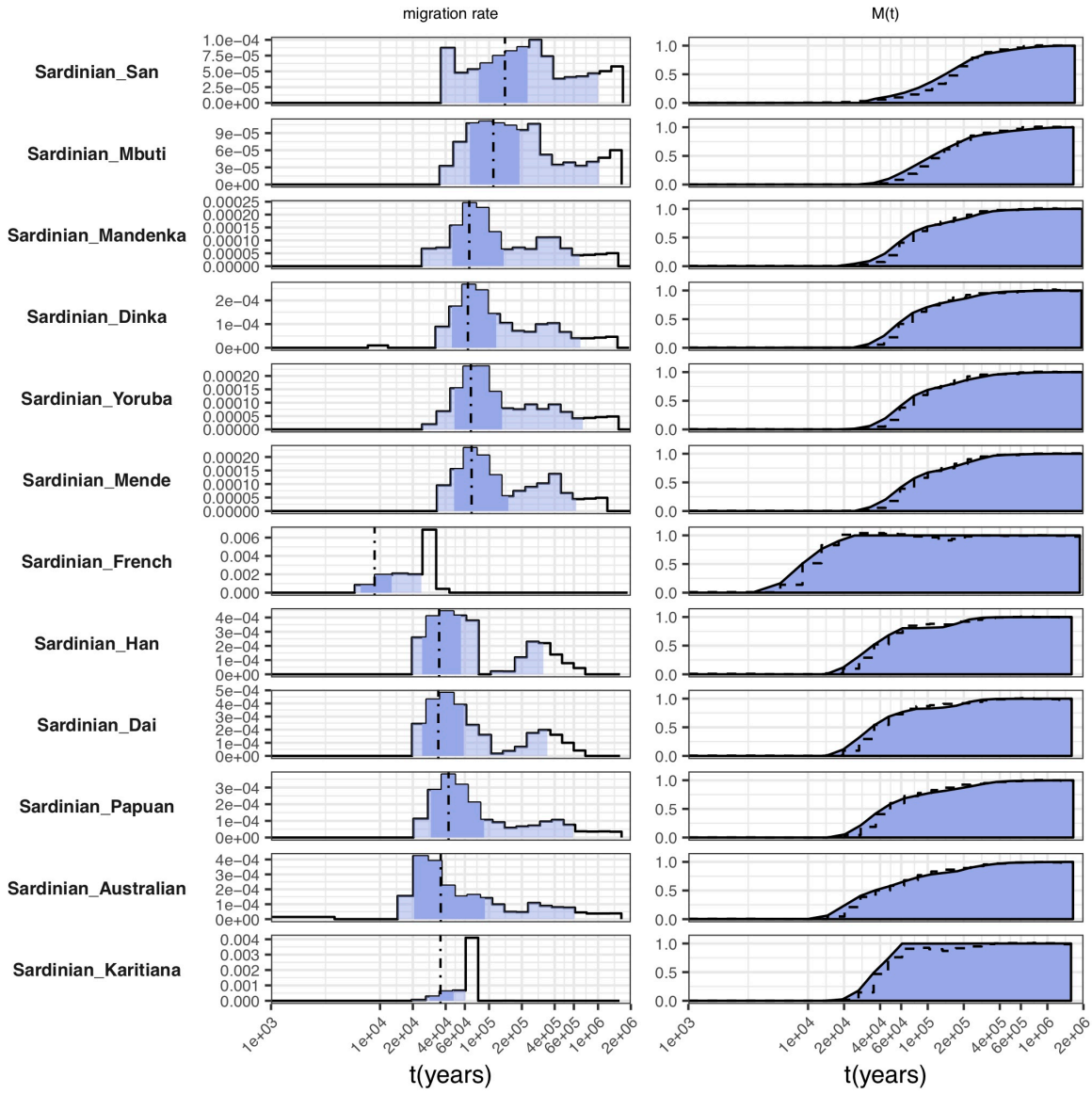
F


migration rate
M(t)
Mende_San
Mende_Mbuti
Mende_Mandenka
Mende_Dinka
Mende_Yoruba
Mende_French
Mende_Sardinian
Mende_Han
Mende_Dai
Mende_Papuan
Mende_Australian
Mende_Karitiana
t(years)
t(years)

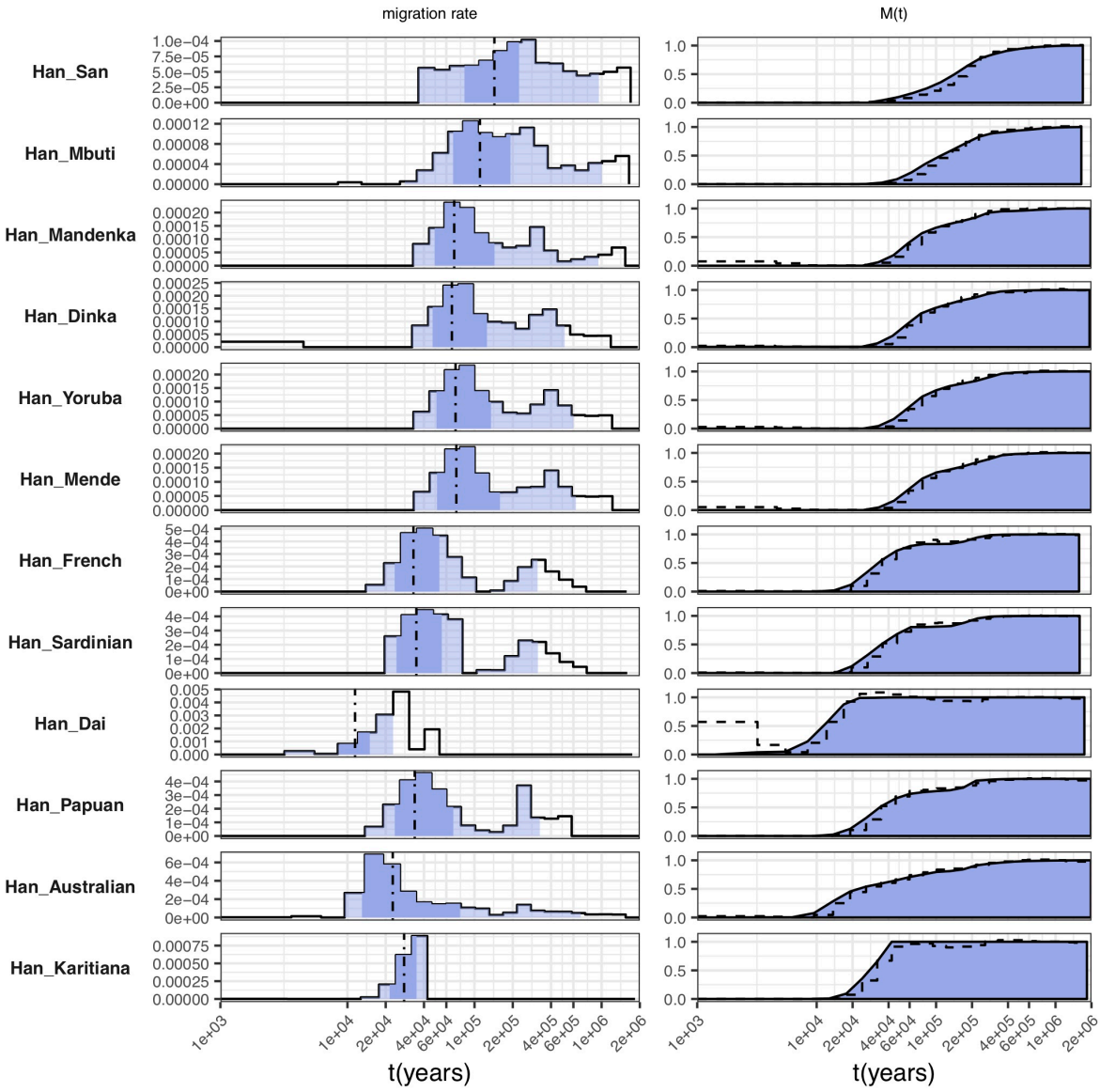
G


migration rate
M(t)
French_San
French_Mbuti
French_Mandenka
French_Dinka
French_Yoruba
French_Mende
French_Sardinian
French_Han
French_Dai
French_Papuan
French_Australian
French_Karitiana
t(years)
t(years)

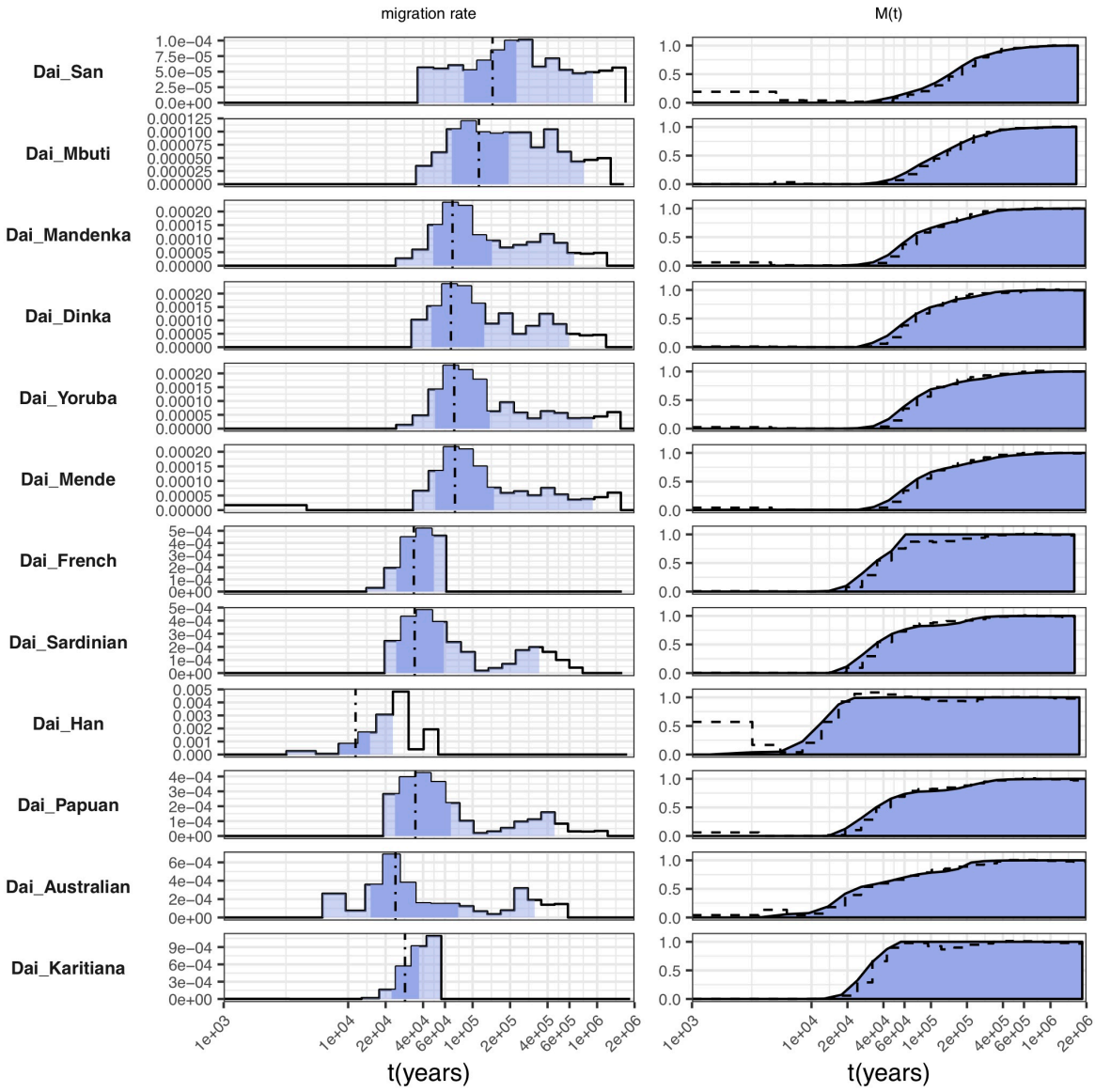
H


migration rate
M(t)
Sardinian_San
Sardinian_Mbuti
Sardinian_Mandenka
Sardinian_Dinka
Sardinian_Yoruba
Sardinian_Mende
Sardinian_French
Sardinian_Han
Sardinian_Dai
Sardinian_Papuan
Sardinian_Australian
Sardinian_Karitiana
1e+03
1e+04
2e+04
4e+04
6e+04
1e+05
2e+05
4e+05
6e+05
1e+06
2e+06
t(years)
t(years)

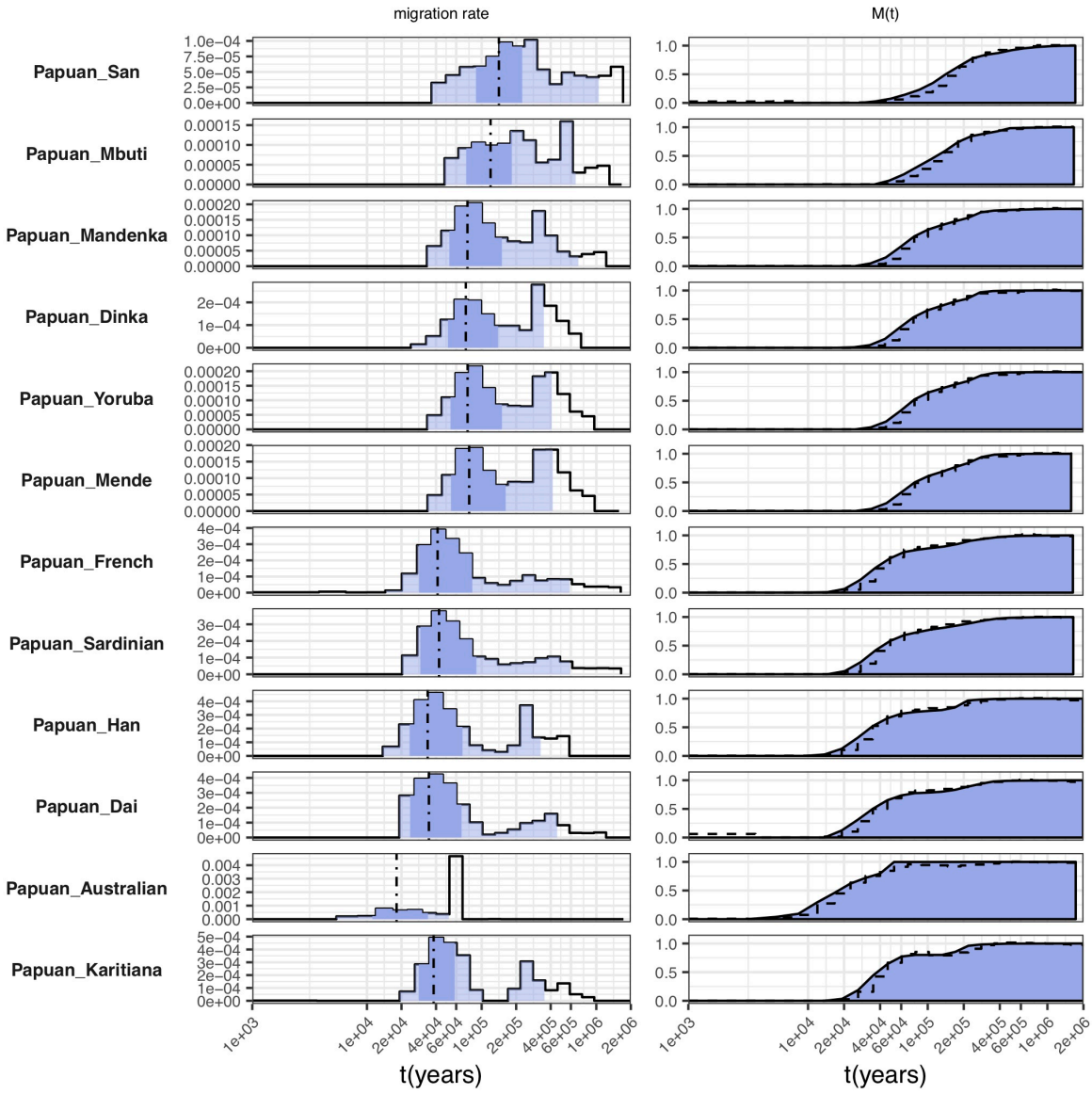
I


migration rate
M(t)
Han_San
Han_Mbuti
Han_Mandenka
Han_Dinka
Han_Yoruba
Han_Mende
Han_French
Han_Sardinian
Han_Dai
Han_Papuan
Han_Australian
Han_Karitiana
1e+03
1e+04
2e+04
4e+04
6e+04
1e+05
2e+05
4e+05
6e+05
1e+06
2e+06
t(years)
t(years)

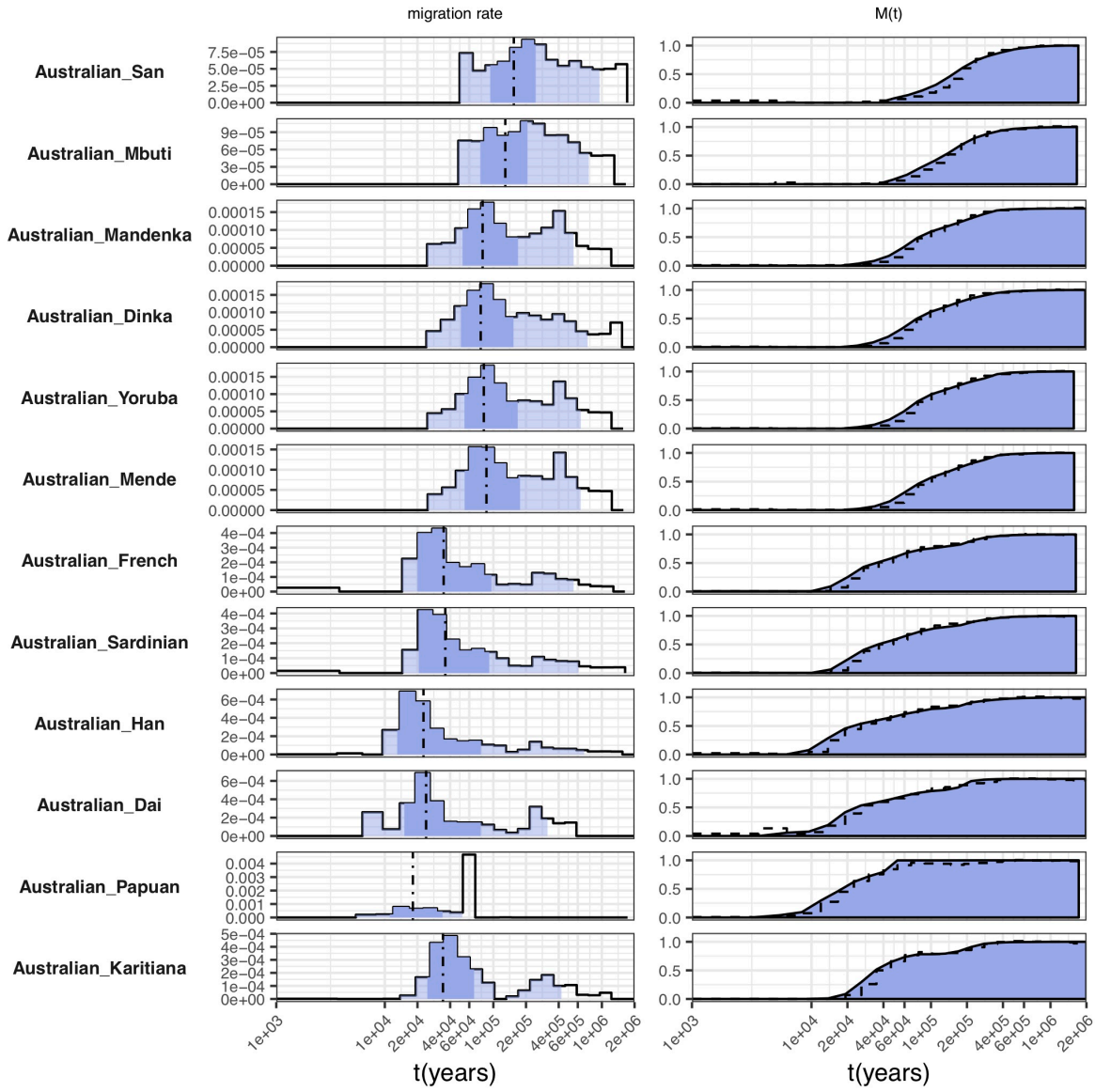
J


migration rate
M(t)
Dai_San
Dai_Mbuti
Dai_Mandenka
Dai_Dinka
Dai_Yoruba
Dai_Mende
Dai_French
Dai_Sardinian
Dai_Han
Dai_Papuan
Dai_Australian
Dai_Karitiana
1e−03
1e+04
2e+04
4e+04
6e+04
1e+05
2e+05
4e+05
6e+05
1e+06
2e+06
t(years)
t(years)

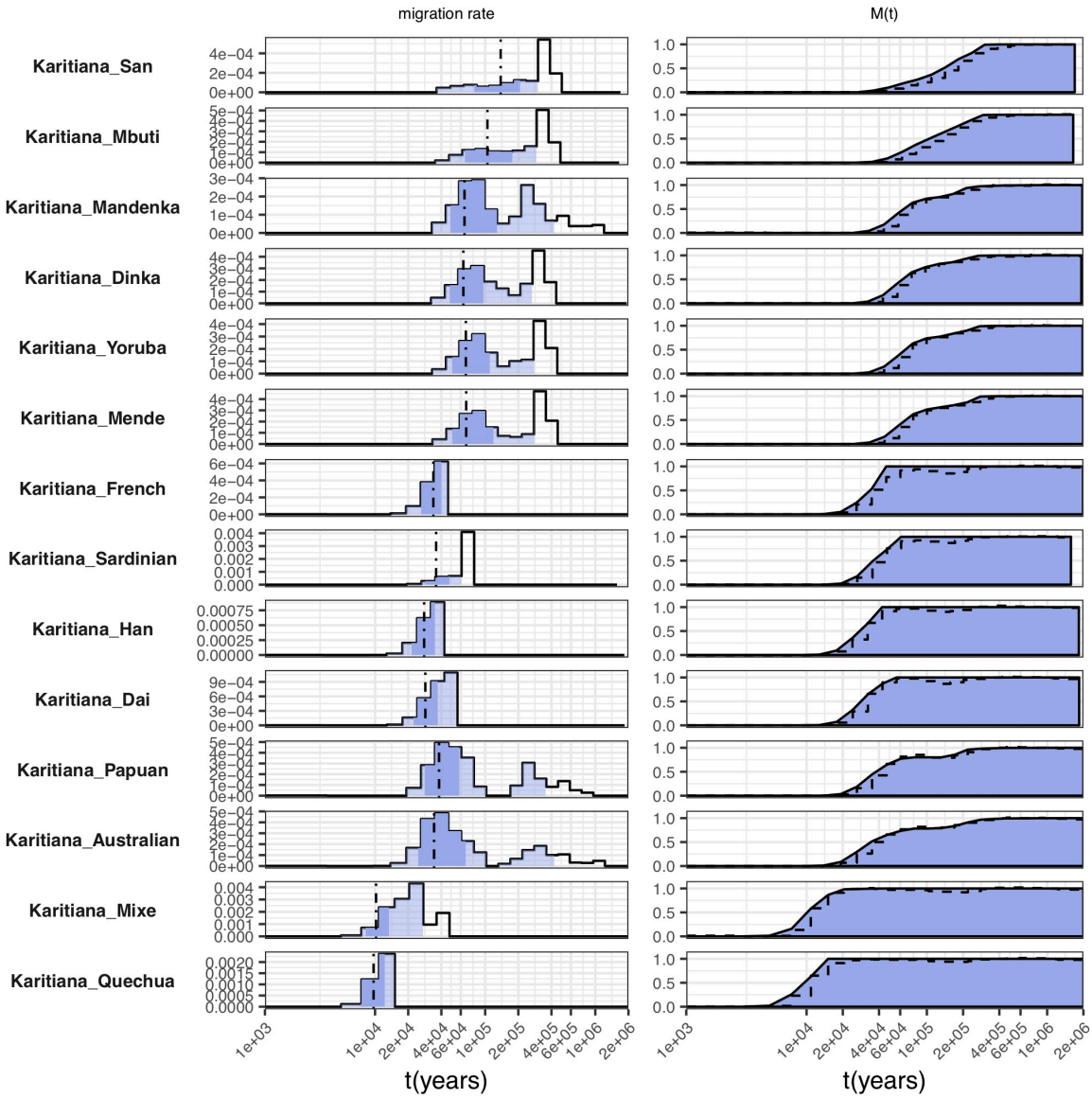
K


migration rate
M(t)
Papuan_San
Papuan_Mbuti
Papuan_Mandenka
Papuan_Dinka
Papuan_Yoruba
Papuan_Mende
Papuan_French
Papuan_Sardinian
Papuan_Han
Papuan_Dai
Papuan_Australian
Papuan_Karitiana
1e+03
1e+04
2e+04
4e+04
6e+04
1e+05
2e+05
4e+05
6e+05
1e+06
2e+06
t(years)
t(years)

L

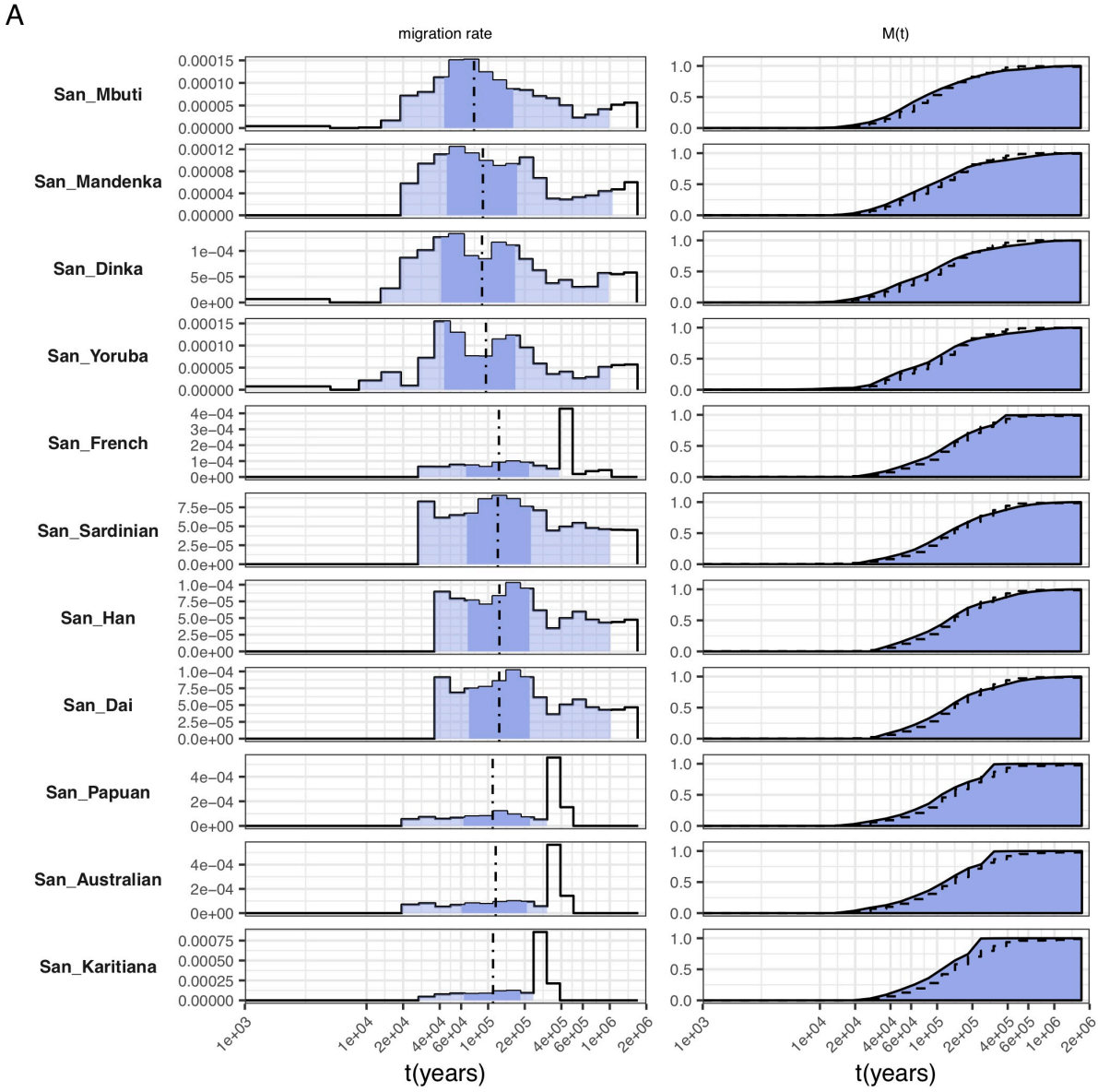

migration rate
M(t)
Australian_San
Australian_Mbuti
Australian_Mandenka
Australian_Dinka
Australian_Yoruba
Australian_Mende
Australian_French
Australian_Sardinian
Australian_Han
Australian_Dai
Australian_Papuan
Australian_Karitiana
t(years)
t(years)

M

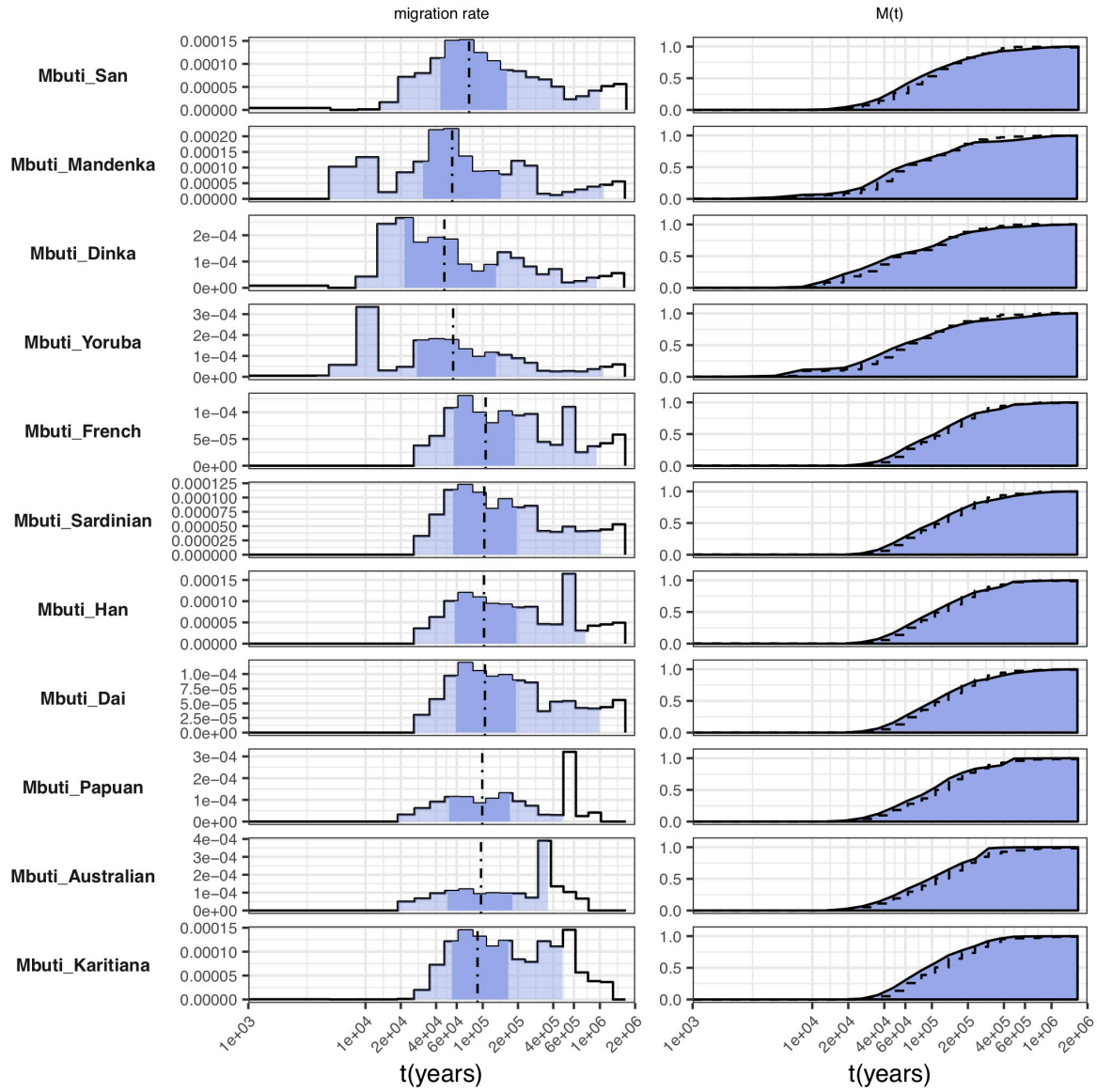


S4 Fig. Pairwise migration profiles for 13 worldwide populations, involving San (A), Mbuti (B), Mandenka (C), Dinka (D), Yoruba (E), Mende (F), French (G), Sardinian (H), Han (I), Dai (J), Papuan (K), Australian (L), Karitiana (M). The relative CCR is shown in step-wise dashed lines to be compared with M(t).

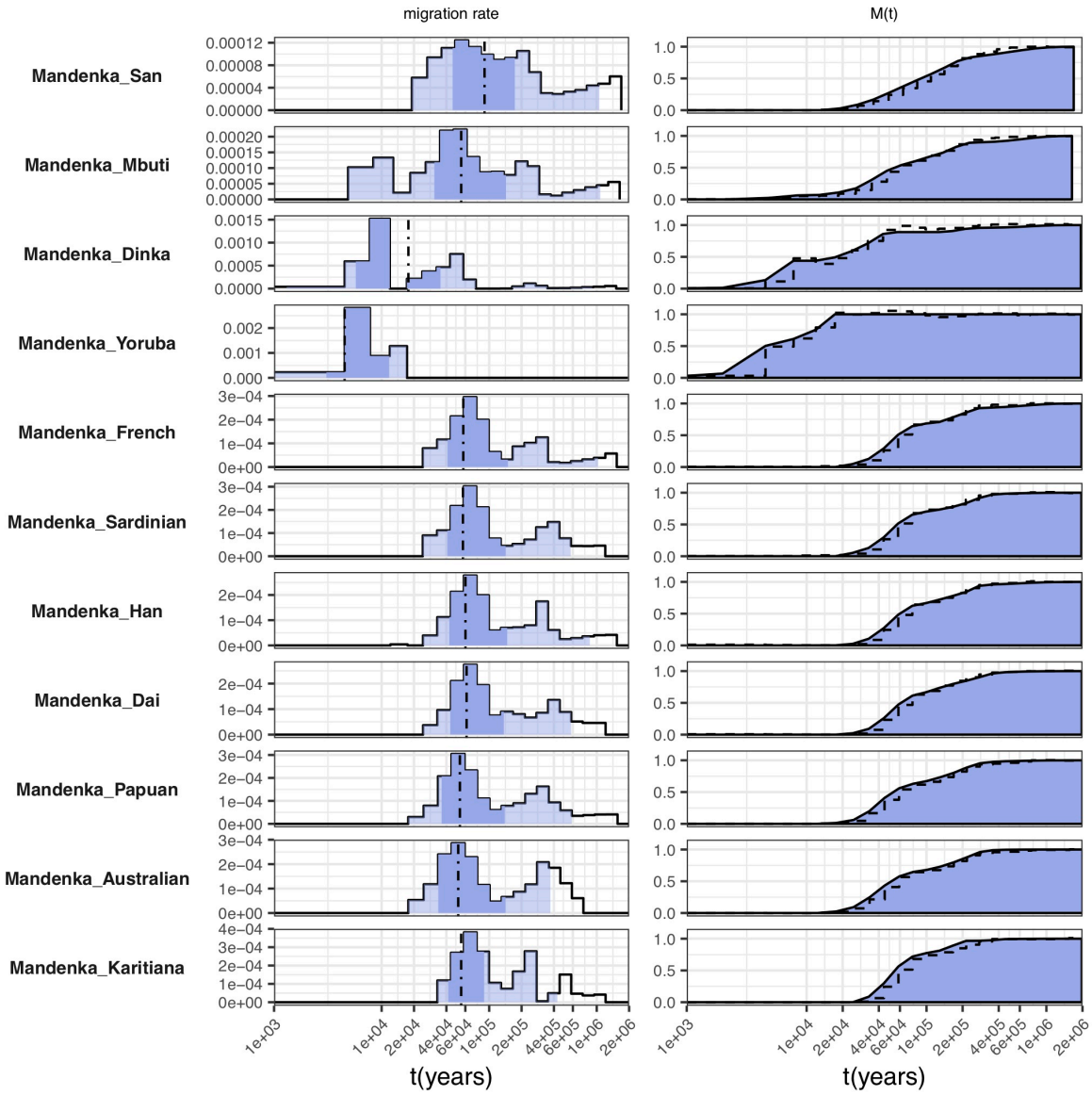
A


migration rate
M(t)
San_Mbuti
San_Mandenka
San_Dinka
San_Yoruba
San_French
San_Sardinian
San_Han
San_Dai
San_Papuan
San_Australian
San_Karitiana
0.00015
0.00010
0.00005
0.00000
0.00012
0.00008
0.00004
1e−04
5e−05
0e+00
4e−04
3e−04
2e−04
7.5e−05
5.0e−05
2.5e−05
0.0e+00
1.0e−04
7.5e−05
5.0e−05
2.5e−05
0.0e+00
0.00075
0.00050
0.00025
1.0
0.5
0.0
1e+03
1e+04
2e+04
4e+04
6e+04
1e+05
2e+05
4e+05
6e+05
1e+06
2e+06
t(years)

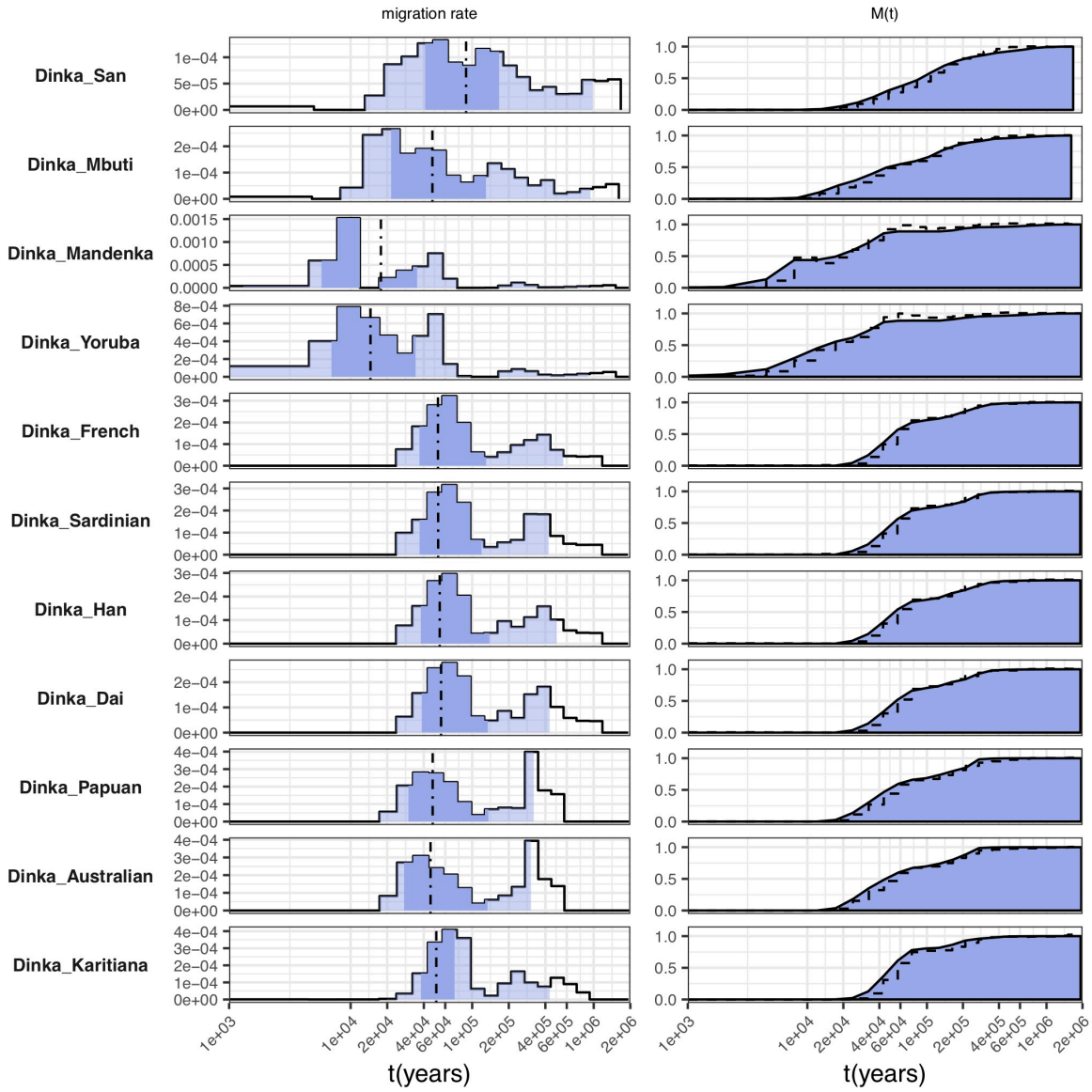
B


migration rate
M(t)
Mbuti_San
Mbuti_Mandenka
Mbuti_Dinka
Mbuti_Yoruba
Mbuti_French
Mbuti_Sardinian
Mbuti_Han
Mbuti_Dai
Mbuti_Papuan
Mbuti_Australian
Mbuti_Karitiana
1e+03
1e+04
2e+04
4e+04
6e+04
1e+05
2e+05
4e+05
6e+05
1e+06
2e+06
t(years)
t(years)

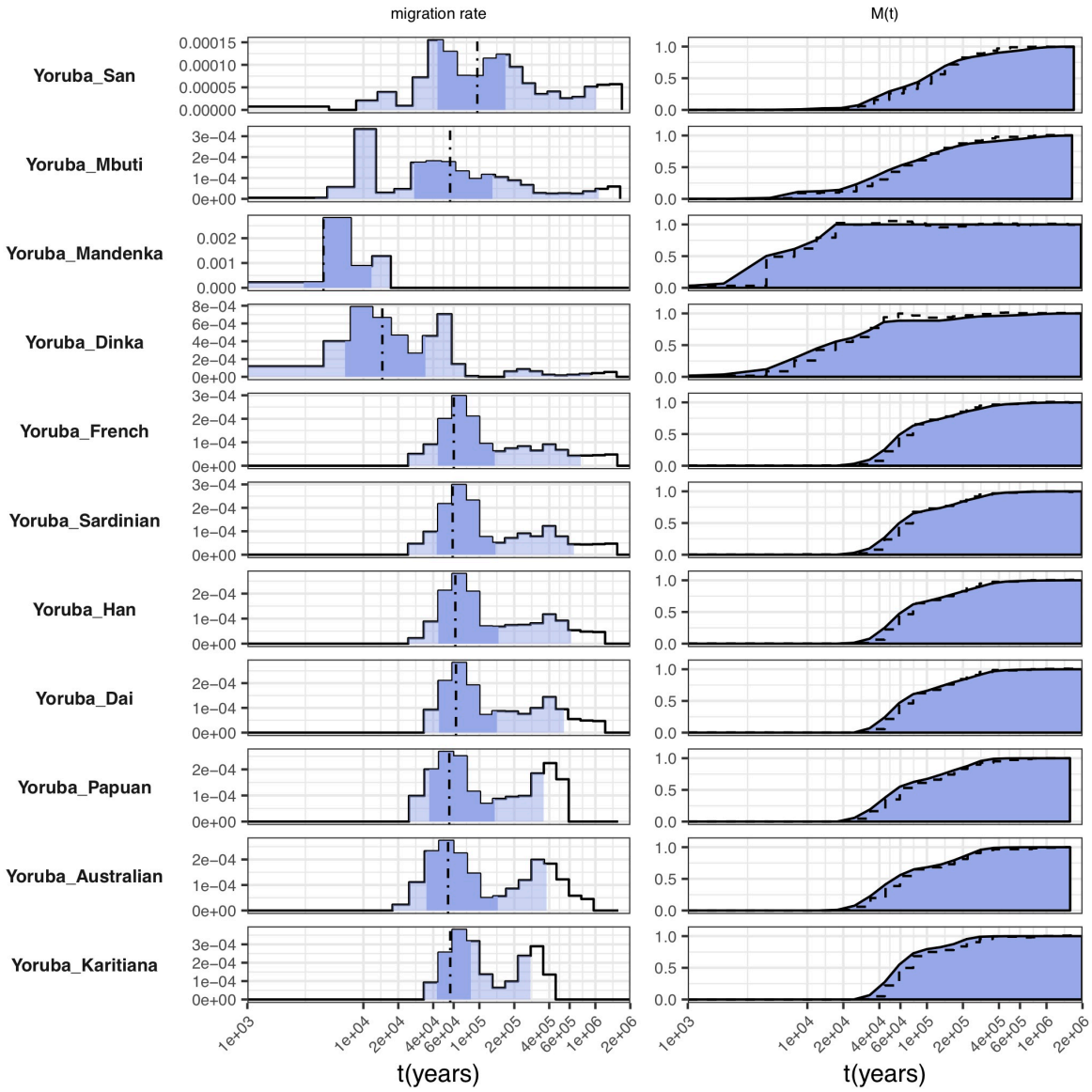
C


migration rate
M(t)
Mandenka_San
Mandenka_Mbuti
Mandenka_Dinka
Mandenka_Yoruba
Mandenka_French
Mandenka_Sardinian
Mandenka_Han
Mandenka_Dai
Mandenka_Papuan
Mandenka_Australian
Mandenka_Karitiana
0.00012
0.00008
0.00004
0.00000
0.00020
0.00015
0.00010
0.00005
0.00000
0.0015
0.0010
0.0005
0.0000
0.002
0.001
0.000
3e−04
2e−04
1e−04
0e+00
4e−04
1.0
0.5
0.0
1e+03
1e+04
2e+04
4e+04
6e+04
1e+05
2e+05
4e+05
6e+05
1e+06
2e+06
t(years)

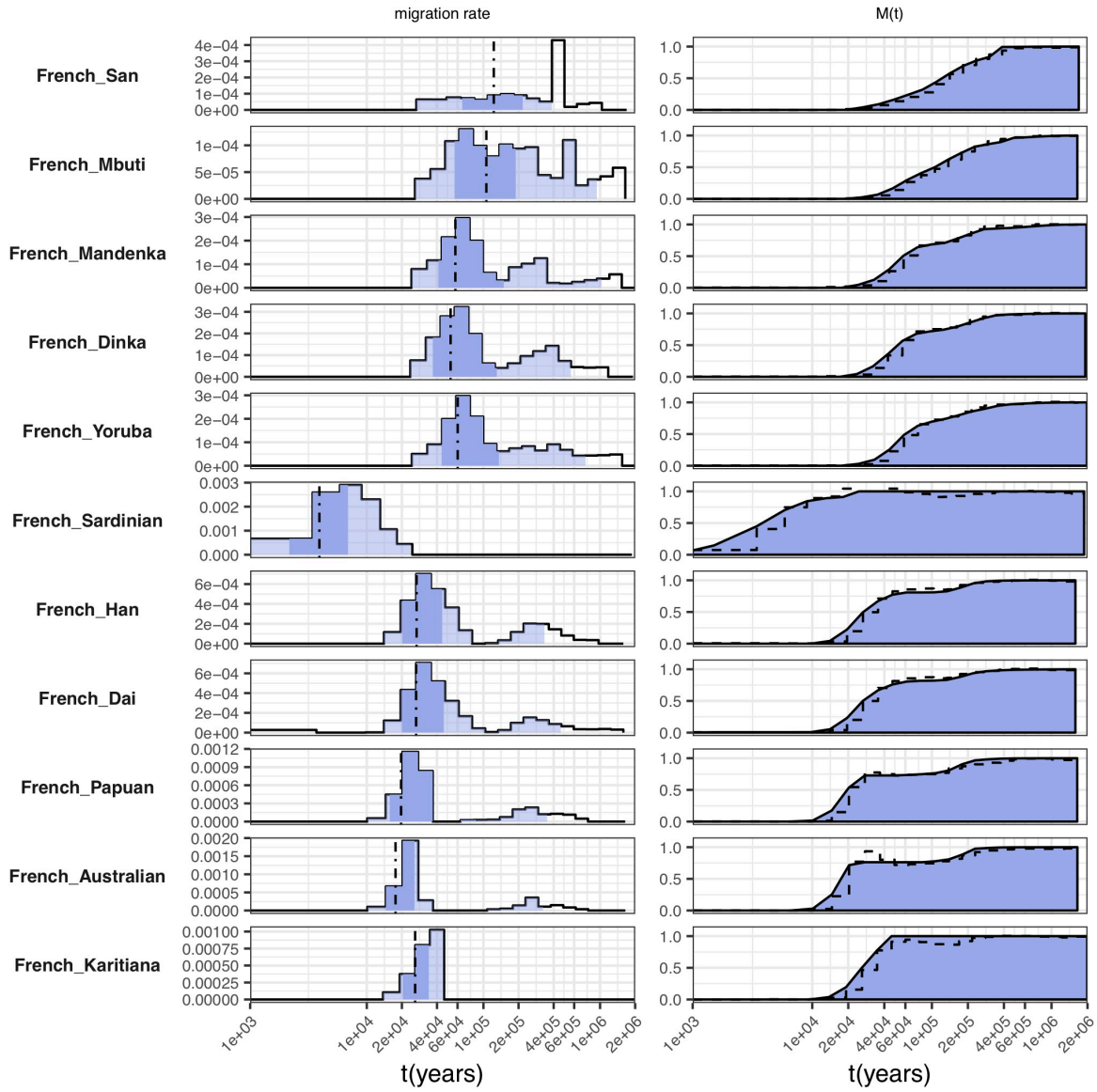
D


migration rate
M(t)
Dinka_San
Dinka_Mbuti
Dinka_Mandenka
Dinka_Yoruba
Dinka_French
Dinka_Sardinian
Dinka_Han
Dinka_Dai
Dinka_Papuan
Dinka_Australian
Dinka_Karitiana
1e−04
5e−05
0e+00
2e−04
1e−04
0e+00
0.0015
0.0010
0.0005
0.0000
8e−04
6e−04
4e−04
2e−04
0e+00
3e−04
2e−04
1e−04
0e+00
4e−04
3e−04
2e−04
1e−04
0e+00
1.0
0.5
0.0
1e+03
1e+04
2e+04
4e+04
6e+04
1e+05
2e+05
4e+05
6e+05
1e+06
2e+06
t(years)

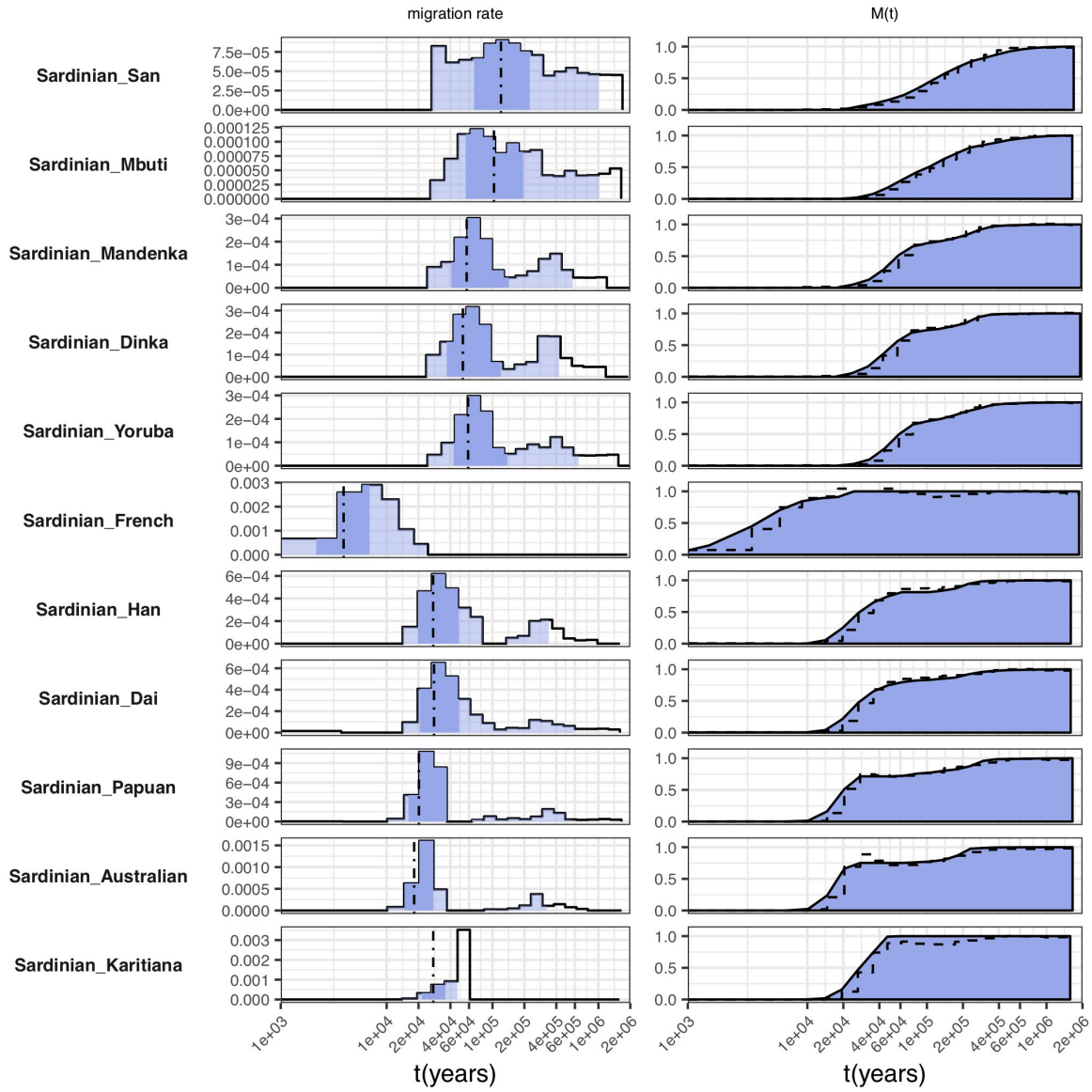
E


migration rate
M(t)
Yoruba_San
Yoruba_Mbuti
Yoruba_Mandenka
Yoruba_Dinka
Yoruba_French
Yoruba_Sardinian
Yoruba_Han
Yoruba_Dai
Yoruba_Papuan
Yoruba_Australian
Yoruba_Karitiana
0.00015
0.00010
0.00005
0.00000
3e−04
2e−04
1e−04
0e+00
0.002
0.001
0.000
8e−04
6e−04
4e−04
2e−04
0e+00
1.0
0.5
0.0
1e+03
1e+04
2e+04
4e+04
6e+04
1e+05
2e+05
4e+05
6e+05
1e+06
2e+06
t(years)

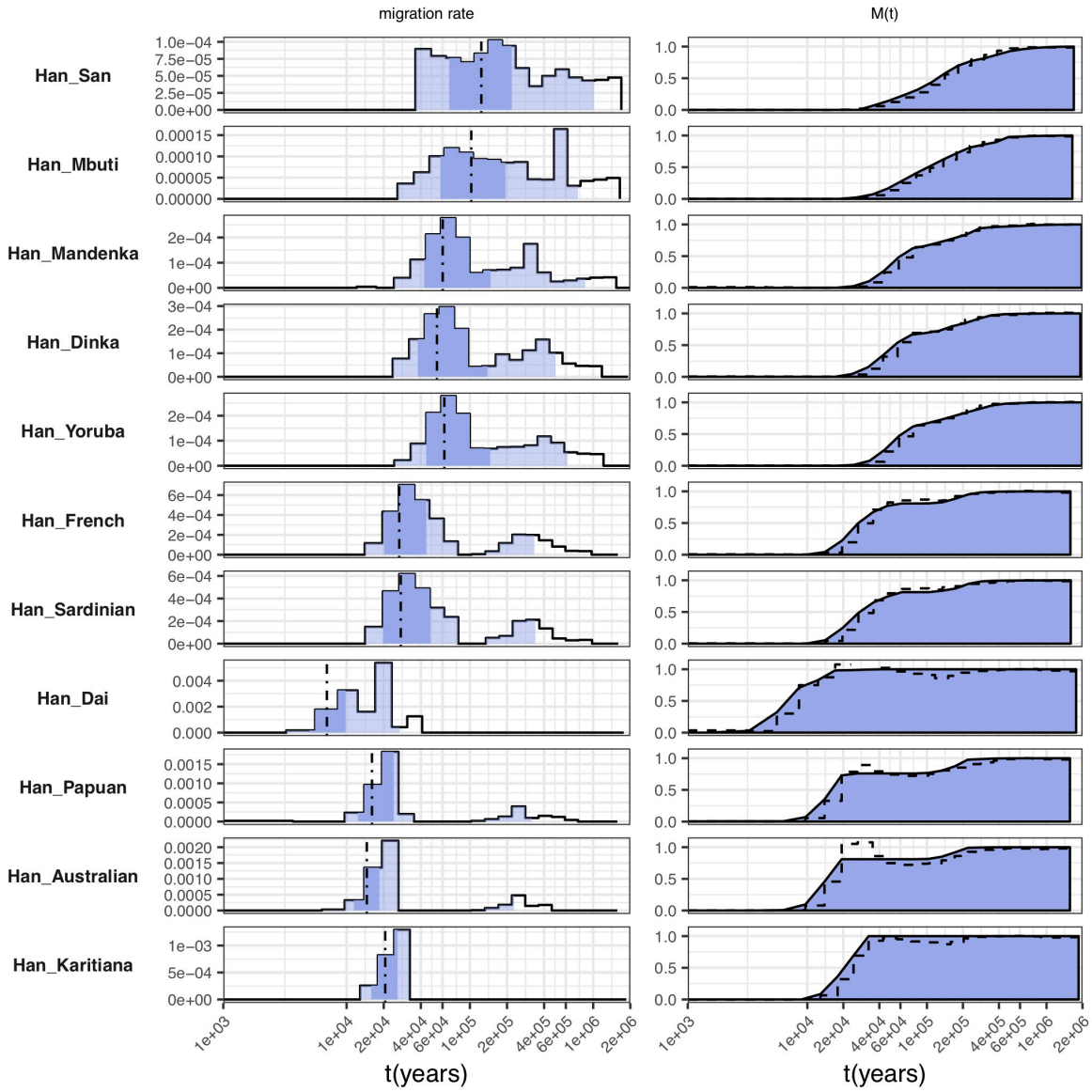
F

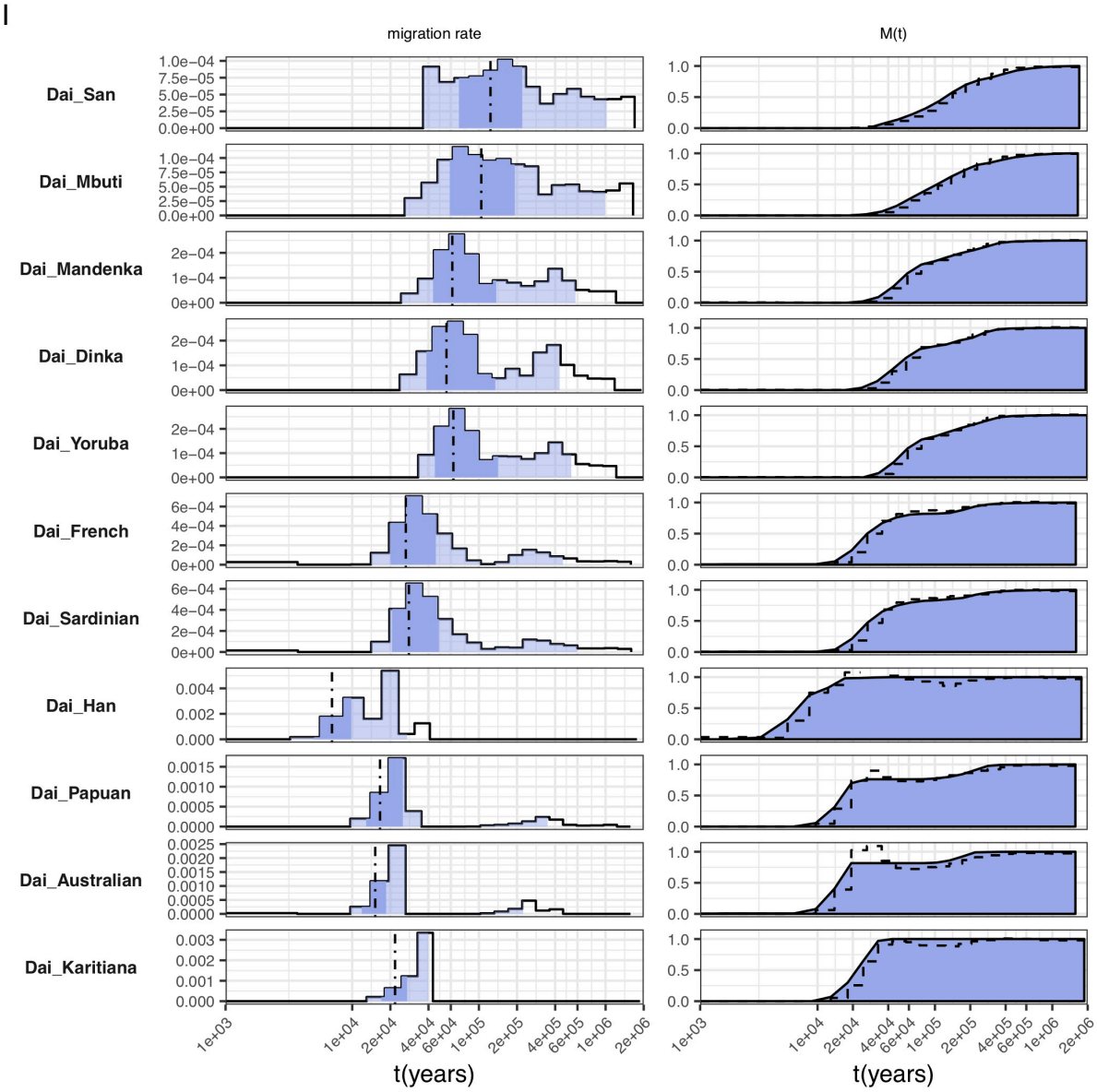

migration rate
M(t)
French_San
French_Mbuti
French_Mandenka
French_Dinka
French_Yoruba
French_Sardinian
French_Han
French_Dai
French_Papuan
French_Australian
French_Karitiana
4e−04
3e−04
2e−04
1e−04
0e+00
5e−05
0.003
0.002
0.001
0.000
6e−04
0.0012
0.0009
0.0006
0.0003
0.0000
0.0020
0.0015
0.0010
0.0005
0.00100
0.00075
0.00050
0.00025
0.00000
1.0
0.5
0.0
1e+03
1e+04
2e+04
4e+04
6e+04
1e+05
2e+05
4e+05
6e+05
1e+06
2e+06
t(years)
t(years)

G

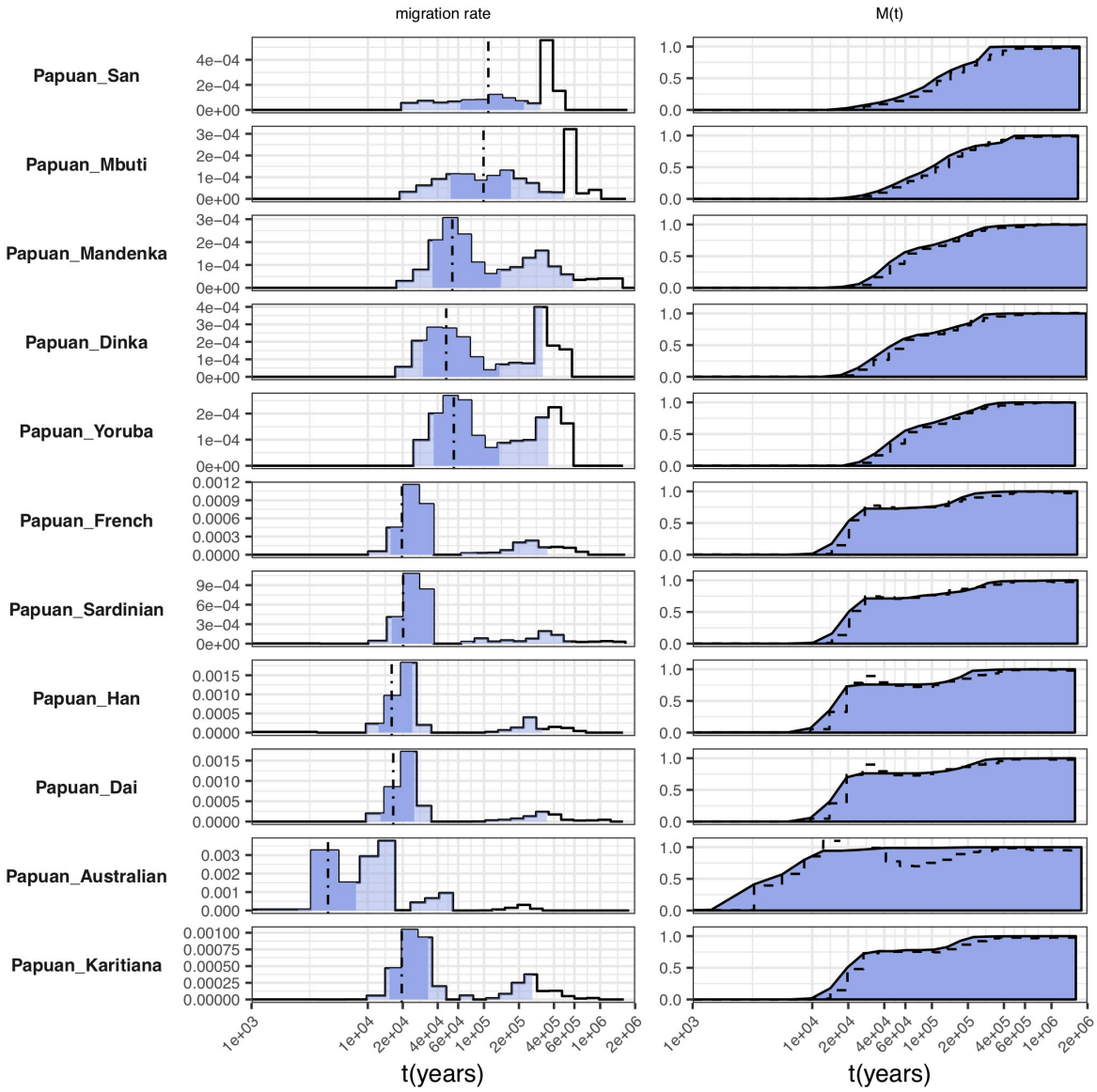

migration rate
M(t)
Sardinian_San
Sardinian_Mbuti
Sardinian_Mandenka
Sardinian_Dinka
Sardinian_Yoruba
Sardinian_French
Sardinian_Han
Sardinian_Dai
Sardinian_Papuan
Sardinian_Australian
Sardinian_Karitiana
7.5e−05
5.0e−05
2.5e−05
0.0e+00
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
3e−04
2e−04
1e−04
0e+00
0.003
0.002
0.001
0.000
6e−04
4e−04
2e−04
9e−04
6e−04
3e−04
0.0015
0.0010
0.0005
0.0000
1.0
0.5
0.0
1e+03
1e+04
2e+04
4e+04
6e+04
1e+05
2e+05
4e+05
6e+05
1e+06
2e+06
t(years)
t(years)

H

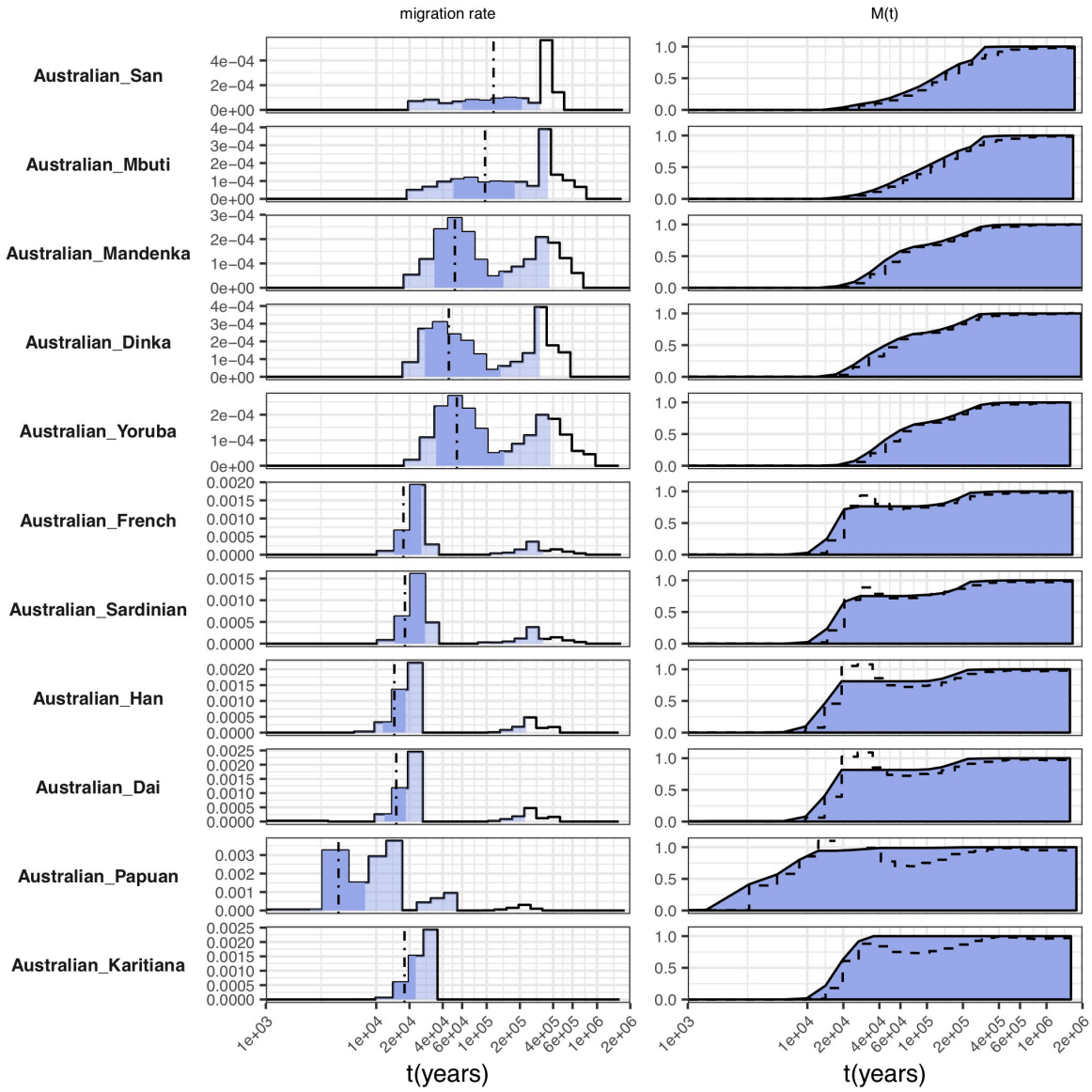

migration rate
M(t)
Han_San
Han_Mbuti
Han_Mandenka
Han_Dinka
Han_Yoruba
Han_French
Han_Sardinian
Han_Dai
Han_Papuan
Han_Australian
Han_Karitiana
1.0e−04
7.5e−05
5.0e−05
2.5e−05
0.0e+00
0.00015
0.00010
0.00005
0.00000
2e−04
1e−04
0e+00
3e−04
2e−04
1e−04
0e+00
2e−04
1e−04
0e+00
6e−04
4e−04
2e−04
0e+00
6e−04
4e−04
2e−04
0e+00
0.004
0.002
0.000
0.0015
0.0010
0.0005
0.0000
0.0020
0.0015
0.0010
0.0005
0.0000
1e−03
5e−04
0e+00
1.0
0.5
0.0
1e+03
1e+04
2e+04
4e+04
6e+04
1e+05
2e+05
4e+05
6e+05
1e+06
2e+06
t(years)
t(years)


migration rate
M(t)
Dai_San
Dai_Mbuti
Dai_Mandenka
Dai_Dinka
Dai_Yoruba
Dai_French
Dai_Sardinian
Dai_Han
Dai_Papuan
Dai_Australian
Dai_Karitiana
1.0e−04
7.5e−05
5.0e−05
2.5e−05
0.0e+00
2e−04
1e−04
0e+00
6e−04
4e−04
2e−04
0.004
0.002
0.000
0.0015
0.0010
0.0005
0.0000
0.0025
0.0020
0.003
0.001
1.0
0.5
0.0
1e+03
1e+04
2e+04
4e+04
6e+04
1e+05
2e+05
4e+05
6e+05
1e+06
2e+06
t(years)
t(years)

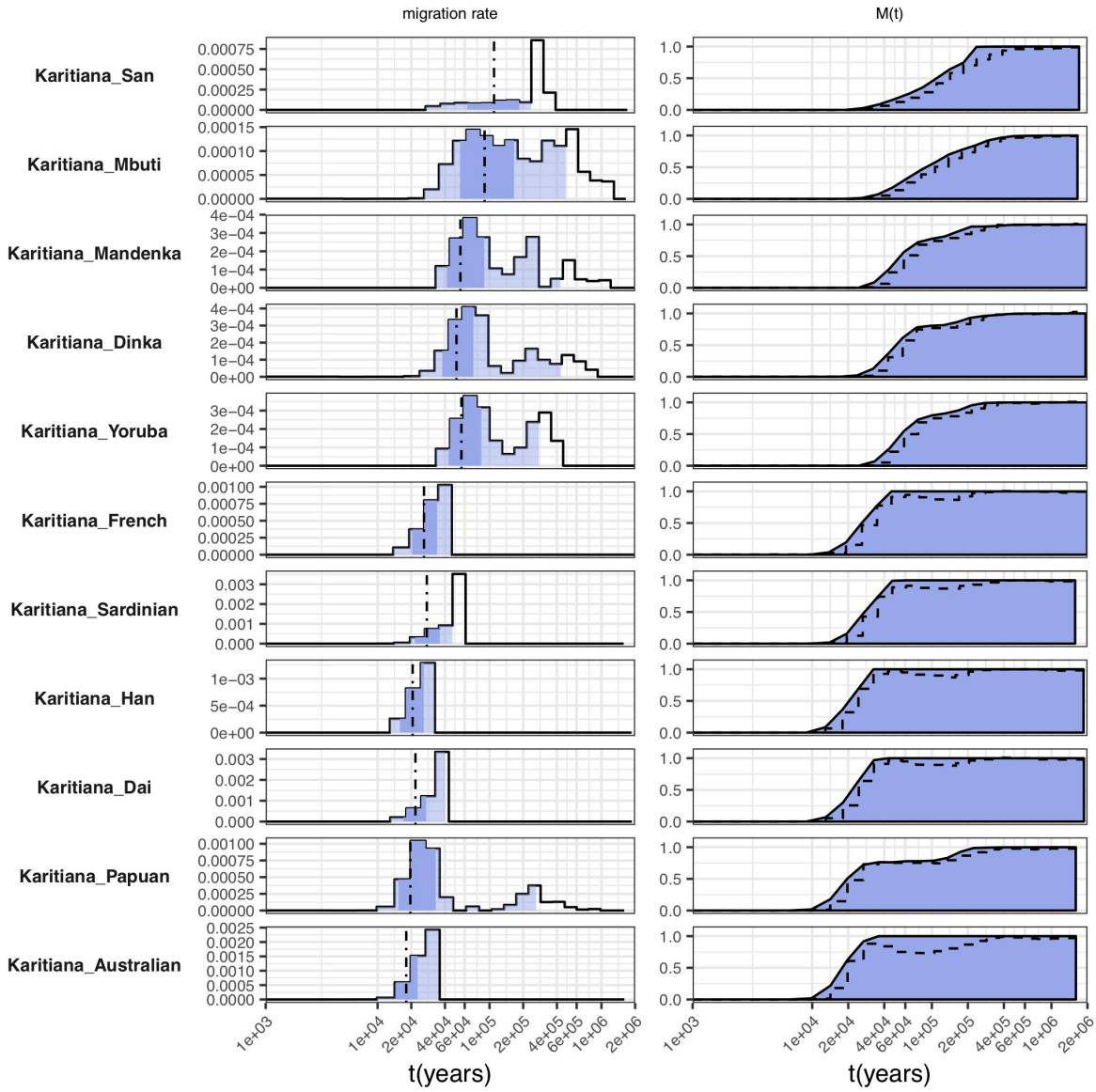
J


migration rate
M(t)
Papuan_San
Papuan_Mbuti
Papuan_Mandenka
Papuan_Dinka
Papuan_Yoruba
Papuan_French
Papuan_Sardinian
Papuan_Han
Papuan_Dai
Papuan_Australian
Papuan_Karitiana
4e−04
2e−04
0e+00
3e−04
1e−04
0.0012
0.0009
0.0006
0.0003
0.0000
9e−04
6e−04
0.0015
0.0010
0.0005
0.003
0.002
0.001
0.000
0.00100
0.00075
0.00050
0.00025
0.00000
1.0
0.5
0.0
1e+03
1e+04
2e+04
4e+04
6e+04
1e+05
2e+05
4e+05
6e+05
1e+06
2e+06
t(years)

K

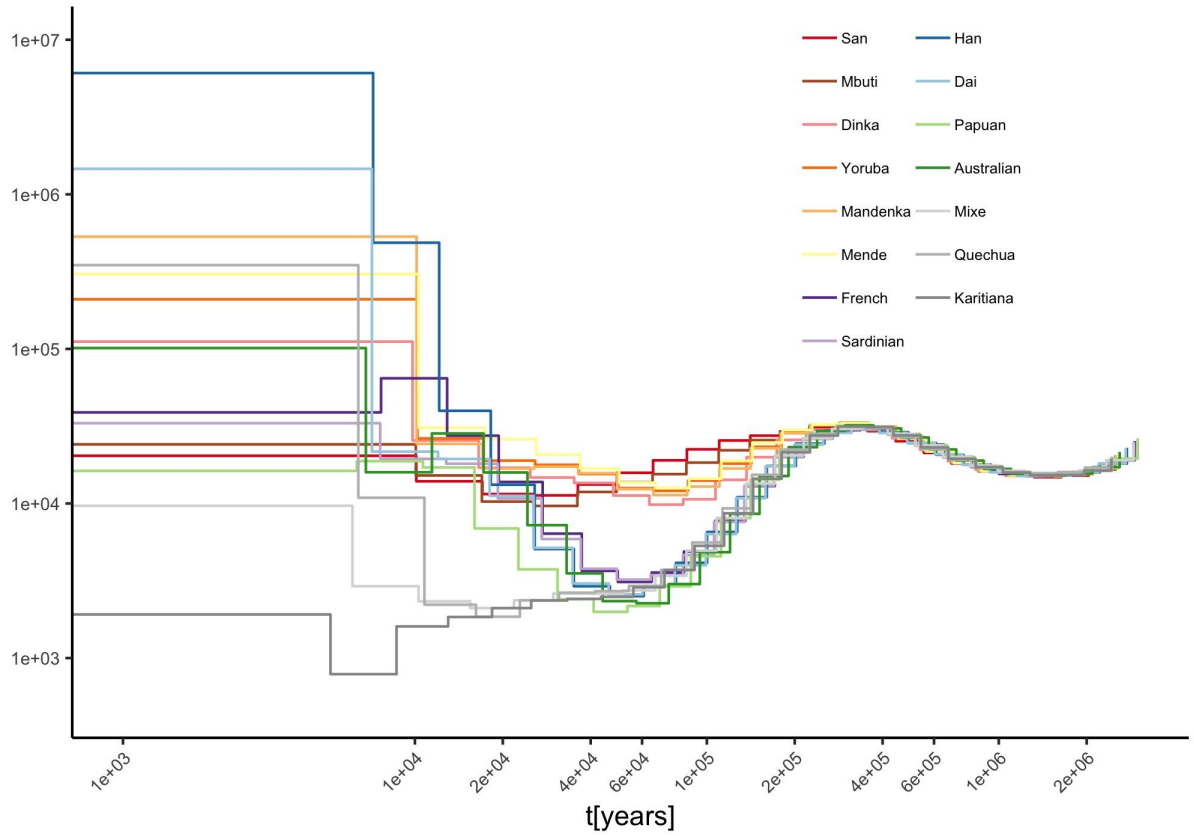

migration rate
M(t)
Australian_San
Australian_Mbuti
Australian_Mandenka
Australian_Dinka
Australian_Yoruba
Australian_French
Australian_Sardinian
Australian_Han
Australian_Dai
Australian_Papuan
Australian_Karitiana
t(years)
t(years)

L

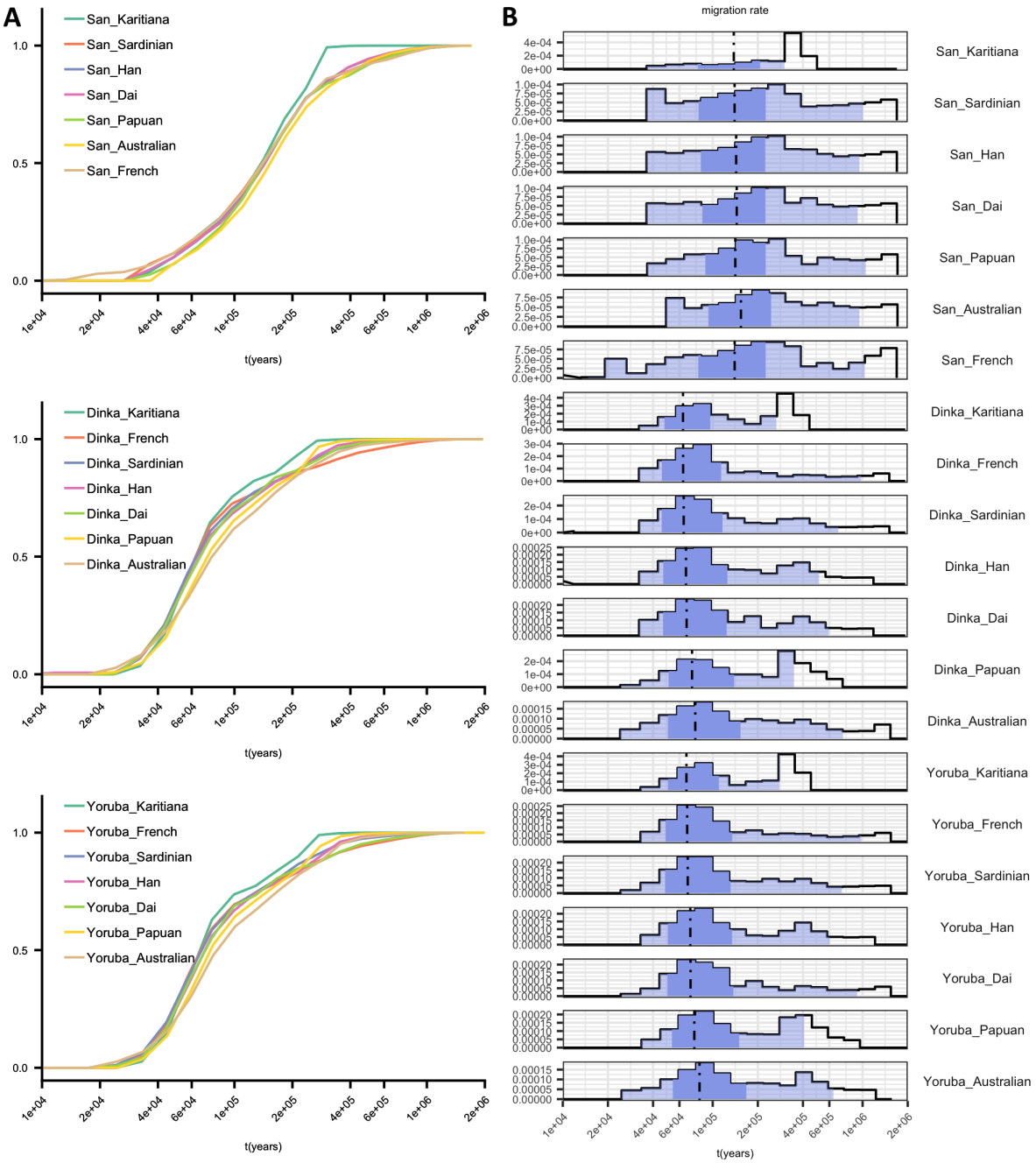


S5 Fig. Migration profile of an independent dataset.

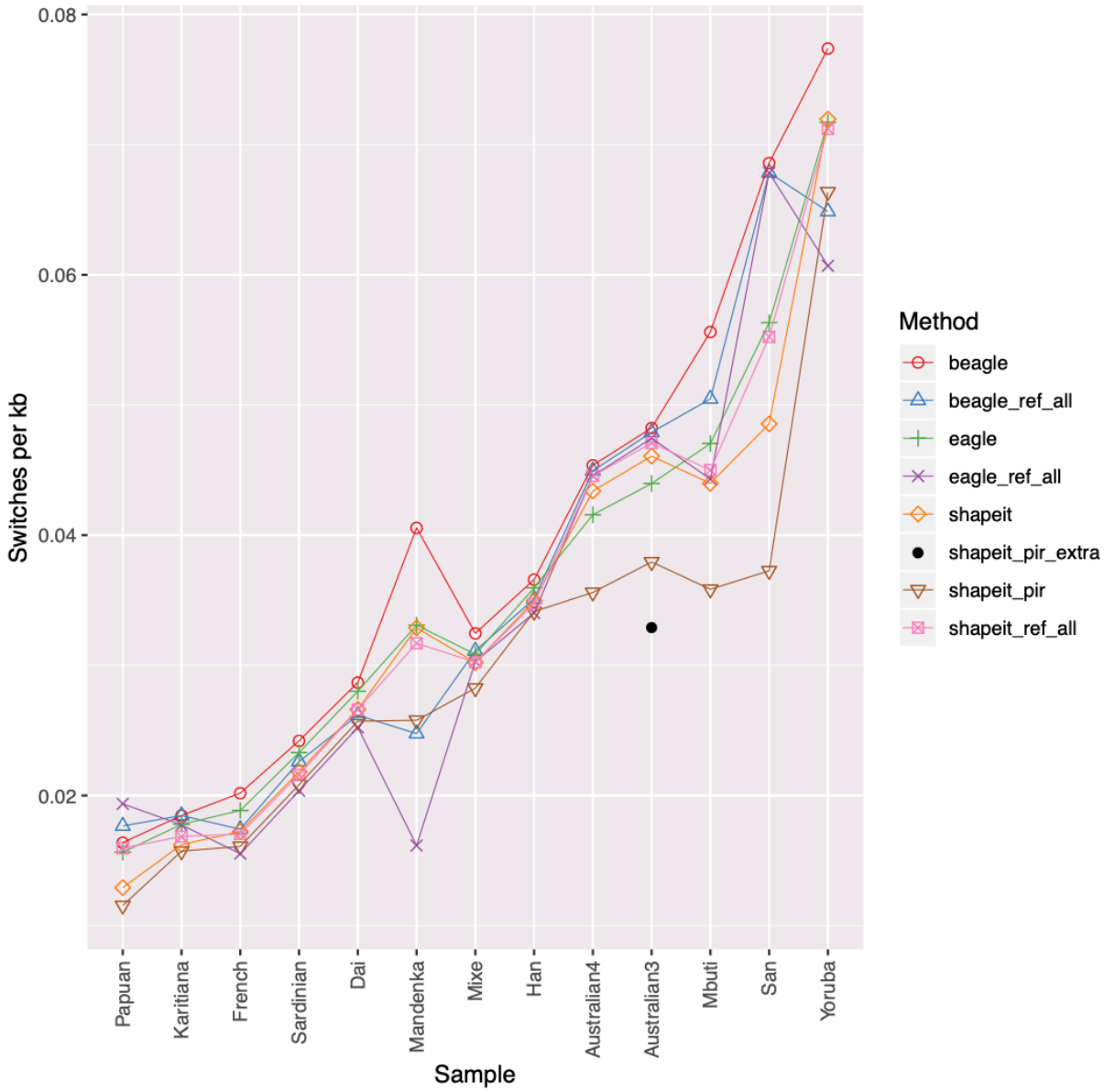
Here we have analysed 12 worldwide populations from Prüfer et al (2014) with independent data processing as described in Methods**:** San (A), Mbuti (B), Mandenka (C), Dinka (D), Yoruba (E), French (F), Sardinian (G), Han (H), Dai (I), Papuan (J), Australian (K), Karitiana (L). The relative CCR is shown in step-wise dashed lines to be compared with M(t).



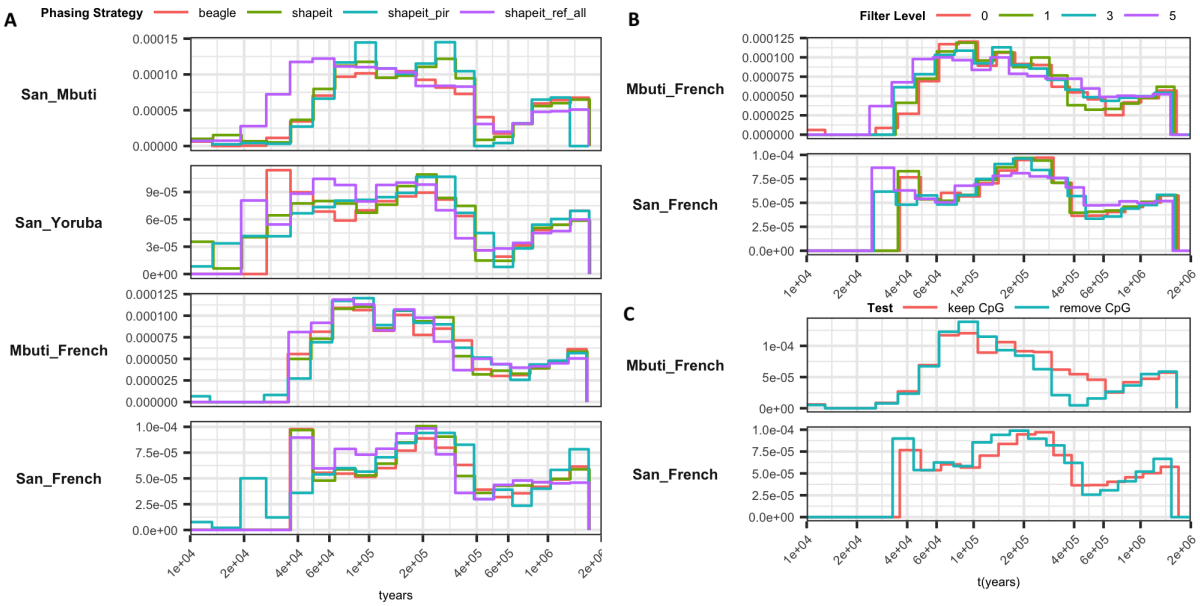
S6 Fig. Estimated population sizes from MSMC2 for 15 worldwide populations. We show the estimates from MSMC using 8 haplotypes/4 individuals per population from the SGDP dateset.



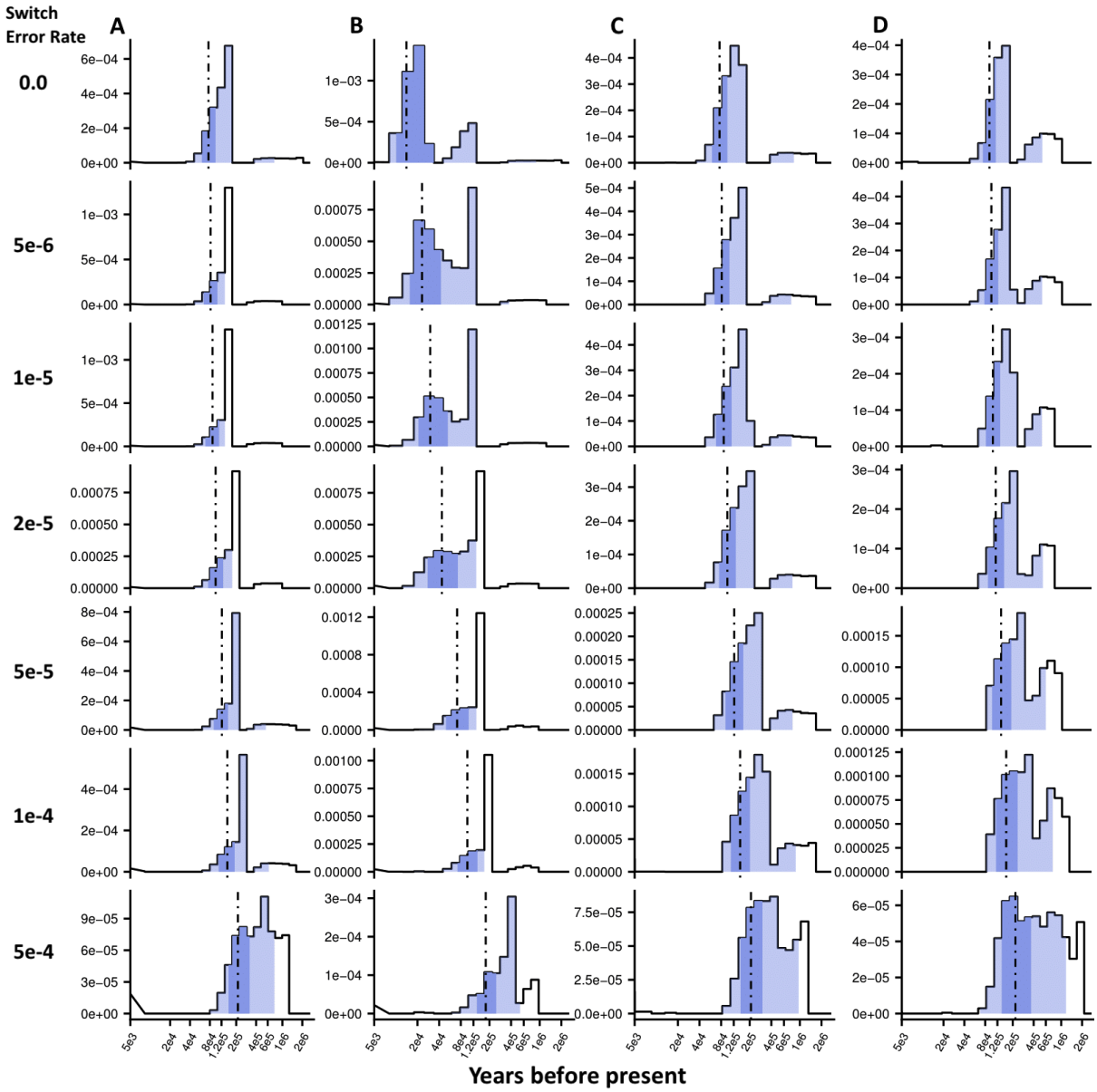
S7 Fig. Testing for potential multiple out-of-Africa separations.
Here we show analyses on the divergence of Papuans and Australians from Africans vs. other Non-African populations from Africans. We show the cumulative migration probability *M(t)* in (A), and the migration rate *m(t)* (B) for pairs of populations of Yoruba, Dinka and San with one non-African population as indicated.



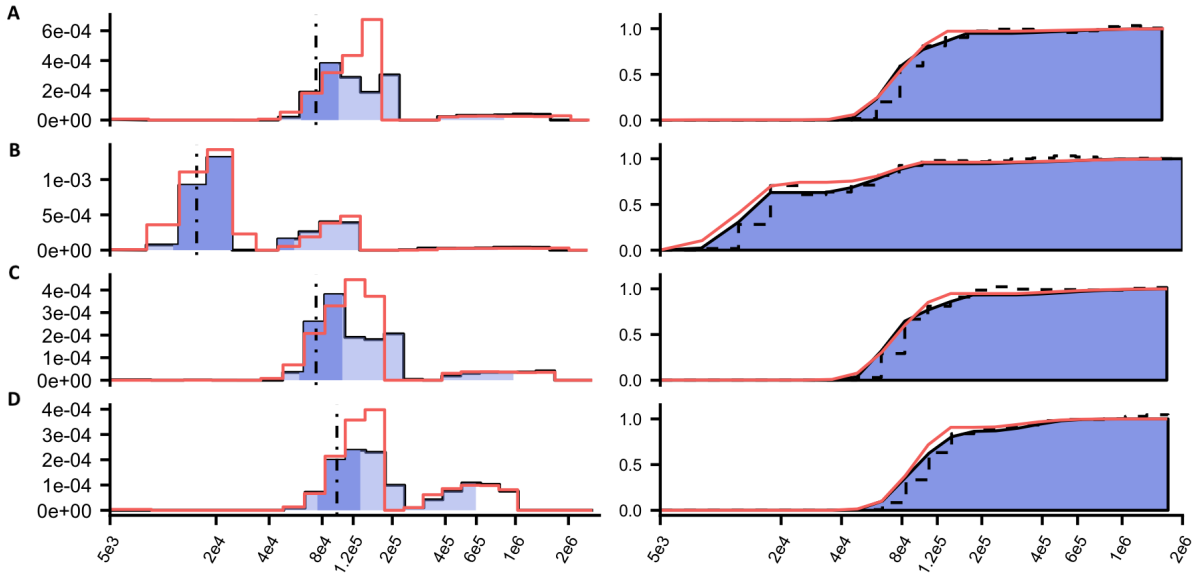
S8 Fig. Switch error rates from eight phasing strategies.
beagle and *beagle_ref_all* denote BEAGLE phasing without and with reference panel (here and below denoting the 1000 Genomes Phase 3 reference panel). *eagle* and *eagle_ref_all* represent EAGLE phasing without and with reference panel. *shapeit* and *shapeit_ref_all* represent SHAPEIT phasing without and with reference panel. *shapeit_pir* represents SHAPEIT phasing with phase-informative reads. *shapeit_pir_extra* represents SHAPEIT phasing with long-insert-size reads as additional phase informative reads, which was applied to B-Australian-3 only. See Methods for details.



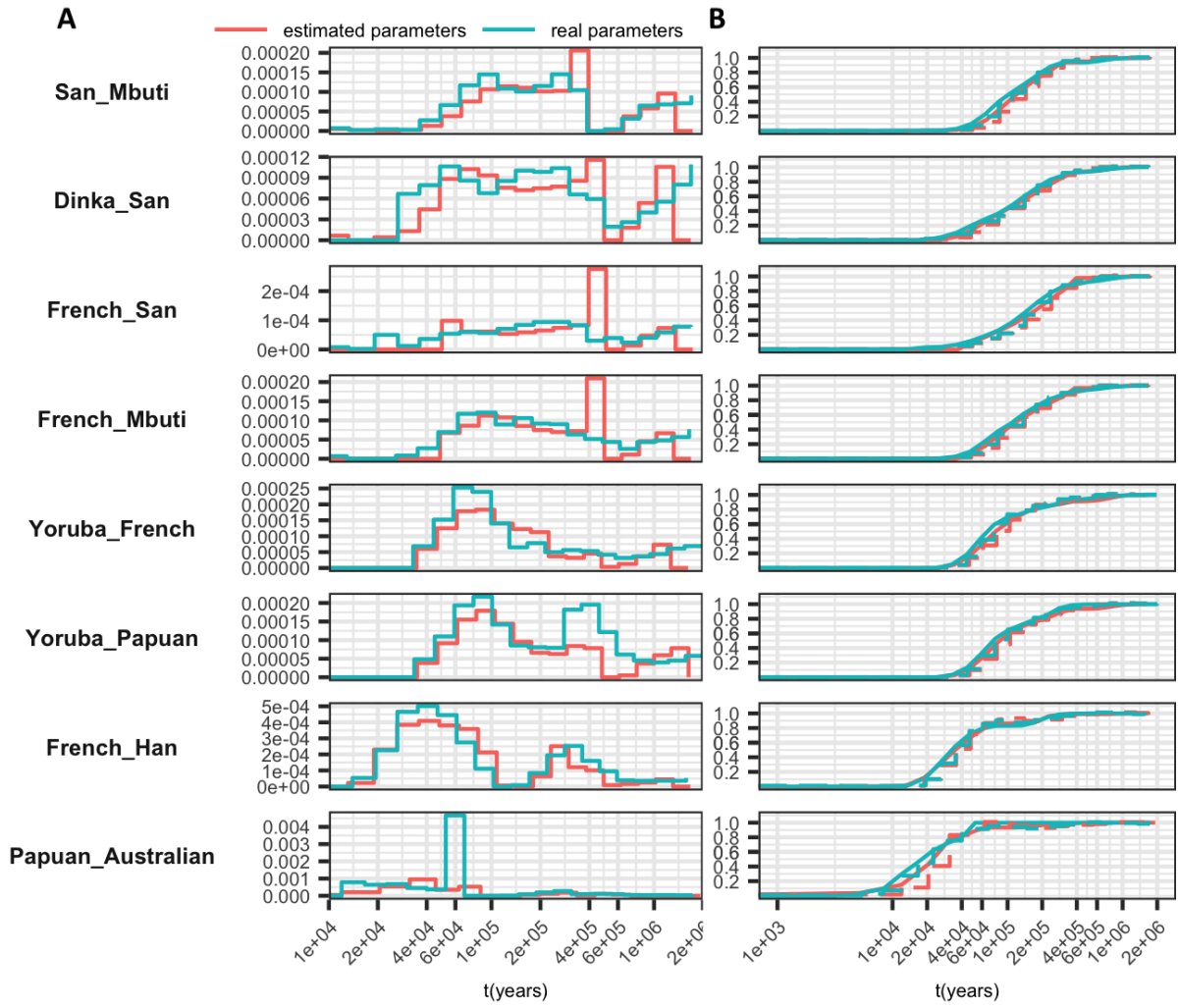
S9 Fig. Impact of phasing and processing artifacts.
We show (A) the impact of the phasing strategy using San/Mbuti, San/Yoruba, Mbuti/French and San/French as examples, (B) the impact of the filtering level for generating individual masks using San/French and Mbuti/French as example, and (C) the impact of removing CpG sites using San/French and Mbuti/French as example. See caption to Figure S8 for a description of the four phasing methods shown in (A).



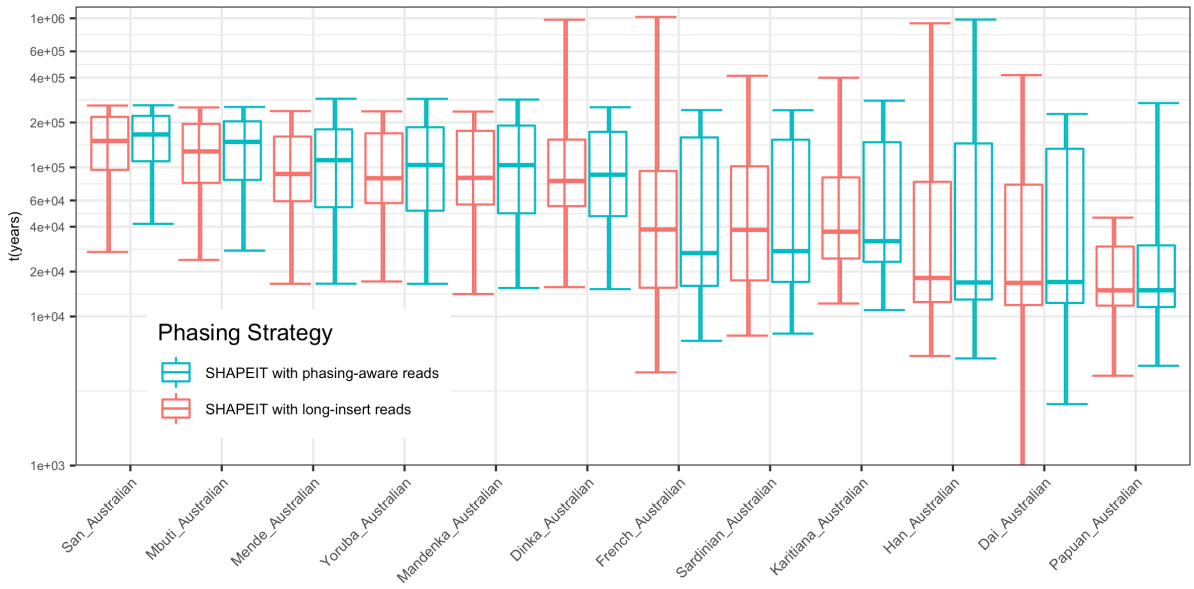
S10 Fig. Impact of switch errors on simulated data.
Here we selected the same four simulation scenarios used in S3 Fig, and added phasing switch errors ranging from 5e-6 to 5e-4 per base pair. The overall migration profiles remain relatively consistent for error rates between 5e-6 and 5e-5, with strong effects seen with rates higher than 5e-5, shifting the migration profiles towards older times. (A) Clean split at 75kya. (B) Split at 75kya with symmetric migration between 10-15kya. (C) Split at 75kya with archaic admixture at 5%. (D) Split at 75kya with archaic admixture at 5% and bottleneck in one population.



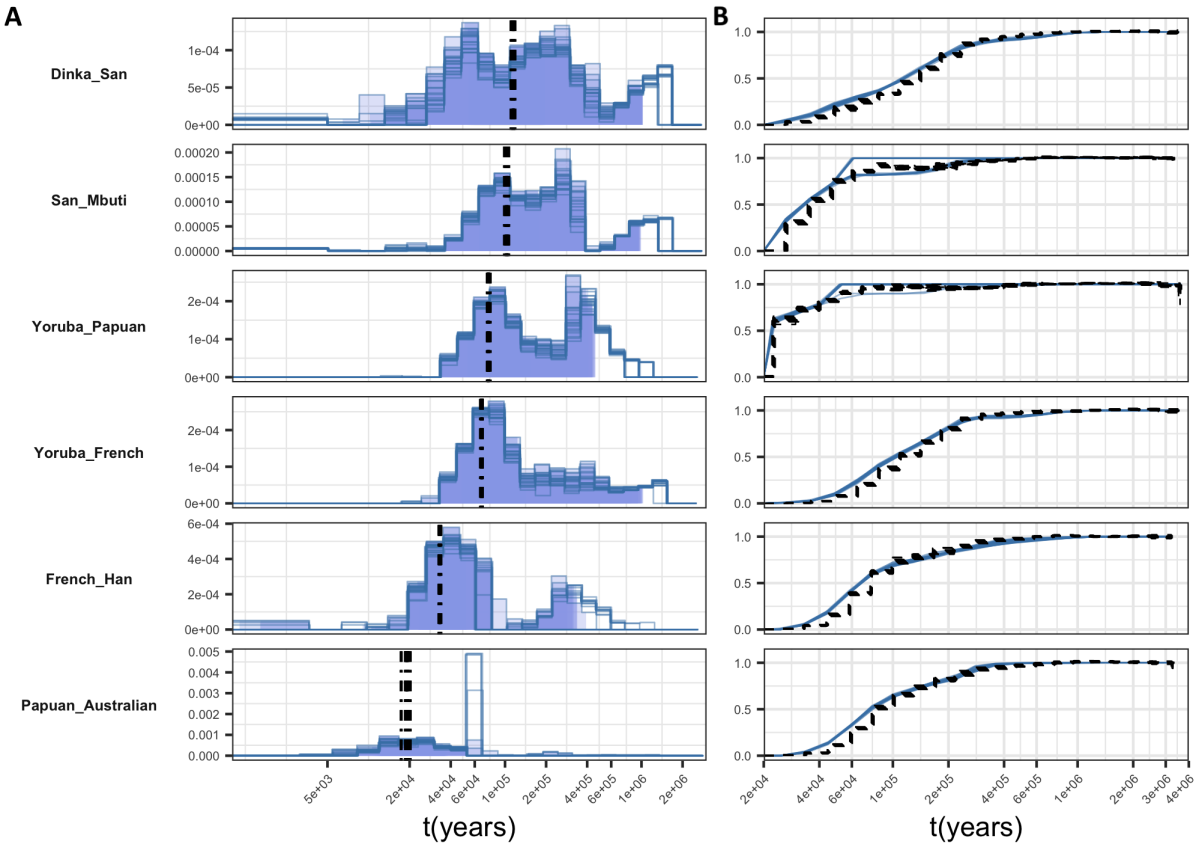
S11 Fig. Impact of recombination rate on simulated data.
Applying the same four simulation scenarios used in S3 Fig, we here used the genetic map estimated for the human genome (i.e. variable recombination rate across genome) instead of a constant recombination rate. Red lines represent our estimates from using a constant recombination rate 10^{-8} per generation per bp. (A) Clean split at 75kya. (B) Split at 75kya with symmetric migration between 10-15kya. (C) Split at 75kya with archaic admixture at 5%. (D) Split at 75kya with archaic admixture at 5% and bottleneck in one population.



S12 Fig. Migration profile on simulated pseudo-SGDP genomes.
Green lines show the estimates we got from SGDP data for pairs shown on the left (as shown in Figure S4), which are used as input parameters for the simulation. Red lines show the estimates from applying MSMC-IM on the simulated data. (A) Migration rates *m(t)*. (B) Cumulative migration probabilities *M(t)* and relative cross-coalescence rates.



S13 Fig. Impact of long-insert phasing on Australian population separation inferences. *M(t)* in quantiles is summarised here between a single Australian and a single individual from worldwide populations**.** Boxes show the 25% to 75% quantiles of *M(t)*, with bi-directional elongated error bars representing 1% and 99% percentiles. Red color represents the data phased using long-insert reads. Green color represents the standard phased dataset.



S14 Fig. Bootstrap tests.
As shown in (A) migration rate m(t) and (B) Cumulative migration probability *M(t)*, the overall inferred profile for each pair is rather consistent across 20 replicates.

S1 Table. Analysed samples and population labels from the SGDP dataset.

Sample ID	Population Label	Country	Continent
S_Yoruba-1	Yoruba	Nigeria	Africa
S_Yoruba-2	Yoruba	Nigeria	Africa
S_Dinka-1	Dinka	Sudan	Africa
S_Dinka-2	Dinka	Sudan	Africa
S_Mbuti-1	Mbuti	Congo	Africa
S_Mbuti-2	Mbuti	Congo	Africa
S_Mandenka-1	Mandenka	Senegal	Africa
S_Mandenka-2	Mandenka	Senegal	Africa
S_Mende-1	Mende	Sierra Leone	Africa
S_Mende-2	Mende	Sierra Leone	Africa
S_Khomani_San-1	San	South Africa	Africa
S_Khomani_San-1	San	South Africa	Africa
S_Sardinian-1	Sardinian	Italy	Europe
S_Sardinian-2	Sardinian	Italy	Europe
S_French-1	French	France	Europe
S_French-2	French	Franch	Europe
S_Han-1	Han	China	Asia
S_Han-2	Han	China	Asia
S_Dai-1	Dai	China	Asia
S_Dai-2	Dai	China	Asia
S_Papuan-1	Papuan	Papua New Guinea	Oceania
S_Papuan-2	Papuan	Papua New Guinea	Oceania
B_Australian-3	Australian	Australia	Australia
B_Australian-4	Australian	Australia	Australia
S_Karitiana-1	Karitiana	Brazil	South America
S_Karitiana-2	Karitiana	Brazil	South America
S_Quechua-1	Quechua	Peru	South America
S_Quechua-2	Quechua	Peru	South America
S_Mixe-2	Mixe	Mexico	South America
S_Mixe-3	Mixe	Mexico	South America

S2 Table. MSMC2 results and MSMC-IM estimates for all pairs of SGDP populations analysed in this paper, *see separate Excel file*. The columns reported are described within a legend included in the Excel file. Downloadable at https://doi.org/10.1371/journal.pgen.1008552.s016 (XLSL)

S1 Text: Derivation of MSMC2 and MSMC-IM theory.

Supplementary Note for the article "Tracking deep human population structure through time from whole genome sequences"

Ke Wang and Stephan Schiffels

1 MSMC2

MSMC, introduced first in [7] was based on a Hidden Markov Model (HMM) to model the first coalescence event in any two haplotypes in multiple individuals. This approach improved resolution in recent time over PSMC, while sacrificing resolution in ancient times. The newer development MSMC2, first implemented and used in [4], uses a model that is simpler than the HMM of MSMC, and at the same time more powerful. The idea is to run a two-haplotype HMM (called PSMC') on all pairs in a set of multiple haplotypes. The likelihood of the entire data is then multiplied as a composite likelihood. The basic PSMC'-HMM uses only pairs of sequences and hence models only a single coalescence time across a pair of sequences. PSMC' is very similar to PSMC ([3]), but more accurately approximates the coalescent with recombination. More specifically, the SMC' [5], which underlies PSMC' is a first-order approximation to the coalescent with recombination, while the SMC [6], which underlies PSMC, is not.

1.1 PSMC

Here we briefly rederive the central equations of the PSMC [3]. In the following, we denote the rate of coalescence by $\lambda(t) = (2N(t))^{-1}$. The transition probability is derived from the SMC model by McVean and Cardin [6]. We consider a *given* recombination event, which takes place at time $u < s$ in either of the two branches $m = \{1, 2\}$. This recombination event causes a "floating" branch which coalesces back onto the other branch at time t. The probability for this is given by the probability that *no* coalescence occurred between u and t times the probability that it coalesces exactly at time t:

$$q(t|s, u, m) = \lambda(t) \exp\left(-\int_u^t \lambda(\nu) d\nu\right) \Theta(t - u) \tag{1}$$

where the Heavyside-function is defined as

$$\Theta(t-u) = \begin{cases} 1 & \text{if } t > u \\ 0 & \text{else} \end{cases} \tag{2}$$

and reflects the fact that the transition probability to switch to time t is zero if $u > t$. We show in the Appendix that this conditional probability is properly normalized, i.e. that $\int_0^\infty q(t|s,u,m)\,dt = 1$ for all given s, u and m.

We need to integrate out the two unknown variables u and m, both with uniform probability. The probability that no recombination occurred in either of the two branches of length s is $\exp(-2rs)$. Together this yields:

$$q(t|s) = e^{-2rs}\delta(t-s) + \left(1-e^{-2rs}\right)\frac{1}{2s}\int_0^s \sum_{k=1}^{2} q(t|s,u,m)du. \tag{3}$$

or

$$q(t|s) = e^{-2rs}\delta(t-s) + \left(1-e^{-2rs}\right)\frac{1}{s}\int_0^{\min(s,t)} \lambda(t)\exp\left(-\int_u^t \lambda(\nu)\,d\nu\right)du. \tag{4}$$

1.2 Including Self-coalescence: PSMC'

Marjoram and Wall [5] realized that there was one particular feature missing from the original SMC formulation. An important rationale behind equation 1 is that the recombining "floating" branch will definitely coalesce with the *other* of the two branches, therefore definitely changing the tMRCA to the new value s. However, it is of course possible, that the floating branch will simply coalesce back onto its own branch, therefore resulting in a recombination event that does *not* change the tMRCA.

In order to extend the model to include this self-coalescence, we again consider the probability that the time switches from s to t, given some recombination time u. We can distinguish two cases: for $t > s$, the transition probability is given by the probability that *no* coalescence occurred with either of the two branches $< t$ *and* no coalescence to the single branch between t and s. For $s < t$, the transition probability is given by the probability of coalescing to the *other* branch, rather than to the self-branch. Finally, we have a third class of recombination events which result in $t = s$, namely if the floating branch coalesces back onto its own branch before s.

The conditional probability then reads

$$\begin{aligned} q(t|s,u,m) = \delta(t-s)\frac{1}{2}\left(1-\exp\left(-2\int_u^t \lambda(\nu)\,d\nu\right)\right) + \\ \begin{cases} \lambda(t)\exp\left(-\int_u^t 2\lambda(\nu)d\nu\right)\Theta(t-u) & \text{for } t \le s \\ \lambda(t)\exp\left(-\int_u^s 2\lambda(\nu)d\nu - \int_s^t \lambda(\nu)\,d\nu\right) & \text{for } t > s. \end{cases} \end{aligned} \tag{5}$$

Again, we show in the Appendix, that this conditional probability is normalized. The full transition probability then reads

$$q(t|s) = \delta(t-s)\left(e^{-2rs} + \left(1-e^{-2rs}\right)\frac{1}{2s}\int_0^t \left(1-\exp\left(-2\int_u^t \lambda(\nu)\,d\nu\right)\right)\,du\right) + \left(1-e^{-2rs}\right)\frac{1}{s}\begin{cases}\int_0^t \lambda(t)\exp\left(-\int_u^t 2\lambda(\nu)d\nu\right)du & \text{for } t \le s\\ \int_0^s \lambda(t)\exp\left(-\int_u^s 2\lambda(\nu)d\nu - \int_s^t \lambda(\nu)\,d\nu\right)du & \text{for } t > s.\end{cases} \quad (6)$$

The equilibrium probability is

$$q_0(t) = \lambda(t)L(0;t) \quad (7)$$

with the integral

$$L\left(t_1;t_2\right) = \exp\left(-\int_{t_1}^{t_2}\lambda(\nu)\,d\nu\right). \quad (8)$$

For later purposes, we introduce some more functions. We rewrite the transition matrix

$$q(t|s) = \delta(t-s)q_1(t) + q_2(t|s) \quad (9)$$

with

$$q_1(t) = e^{-2rt} + \left(1-e^{-2rt}\right)\frac{1}{2s}\int_0^t \left(1-L(u;t)^2\right)\,du \quad (10)$$

$$q_2(t|s)|_{t<s} = \left(1-e^{-2rs}\right)\frac{1}{s}\lambda(t)\int_0^t L(u;t)^2 du, \quad (11)$$

$$q_2(t|s)|_{t>s} = \left(1-e^{-2rs}\right)\frac{1}{s}\lambda(t)L(s;t)\int_0^s L(u;s)^2 du. \quad (12)$$

1.2.1 Discrete time intervals

We divide time into a set of n_T intervals that span the entire space from 0 to ∞. In practice, as interval boundaries we use the same boundaries as chosen by PSMC [3], defined as:

$$T_i = \alpha\exp\left(\frac{i}{N_T}\log\left(1+\frac{T_{\max}}{\alpha}\right)-1\right) \quad (13)$$

Here, α and $T_{\max}$ are constants that in the case of PSMC were chosen to be $\alpha = 0.1$ and $t_{\max} = 15$. Note that by construction we have $T_0 = 0$ and $T_{N_T} = \infty$.

This patterning sets of with time patterns approximately linearly distributed through time, and then crosses over to a patterning that is uniformly distributed in log-space. This ensures higher resolution in recent than in ancient times.

For MSMC2, we would like to increase resolution in recent times depending on the number of individuals, i.e. haplotypes we use. For example, with four haplotypes, in recent times we have approximately 6 times more recent coalescent events to analyse compared to just two haplotypes . This should therefore allow us to increase resolution in recent times by 6 fold. We generally set the parameter α in equation 13 to be

$$\alpha = \frac{0.1}{n_{\text{pairs}}} \tag{14}$$

where n_{pairs} is the number of total haplotype pairs analysed. With phased data, and n_{hap} haplotypes from the same population, we have

$$n_{\text{pairs}} = \frac{n_{\text{hap}}(n_{\text{hap}} - 1)}{2} \tag{15}$$

but this can be different if multiple populations or unphased data is anlaysed. For example, if we have four diploid individuals in total, separated evenly into two populations, then we consider all pairs of haplotypes across the two populations, so we have $n_{\text{pairs}} = 16$. If eight diploid individuals from the same population are analysed, and no phasing is available, then we have $n_{\text{pairs}} = 8$. In the MSMC2-implementation, this behaviour can be controlled with the `--pairIndices` flag (see `https://github.com/stschiff/msmc2`). The scaling of α, according to equation 14, is then set automatically by the number of specified pairs.

1.3 Piecewise constant Population sizes

We then define piecewise constant population sizes which correspond to piecewise constant coalescence rates:

$$\lambda(t) = \lambda_\alpha \text{ for } T_\alpha \leq t < T_{\alpha+1}. \tag{16}$$

We now can compute the integral $L\left(t_1; t_2\right)$. Let the next *lower* time boundary from t_1 be β, and the next *lower* time boundary from t_2 be α. We also define $\Delta_\alpha = T_{\alpha+1} - T_\alpha$:

$$L\left(t_1; t_2\right)|_{\alpha \neq \beta} = \exp\left(-\left(T_{\beta+1} - t_1\right)\lambda_\beta - \sum_{\kappa=\beta+1}^{\alpha-1} \lambda_\kappa \Delta_\kappa - \left(t_2 - T_\alpha\right)\lambda_\alpha \right). \tag{17}$$

$$L\left(t_1; t_2\right)|_{\alpha = \beta} = \exp\left(-\left(t_2 - t_1\right)\lambda_\alpha\right). \tag{18}$$

In the following, we denote the next lower index of a given time in the function parameters, with $q_0(t;\alpha)$ meaning that $T_\alpha < t < T_{\alpha+1}$:

$$q_0(t;\alpha) = \lambda_\alpha L(0;t) \tag{19}$$

$$q_1(t;\alpha) = e^{-2rt} + \left(1 - e^{-2rt}\right)\frac{1}{2t}\int_0^t \left(1 - L(u;t)^2\right)\, du \tag{20}$$

For the off-diagonal integrals we first get for $t < s$:

$$q_2(t;\alpha|s)|_{t<s} = \left(1 - e^{-2rs}\right) \frac{1}{s} \lambda(t) \int_0^t L(u;t)^2 du \tag{21}$$

For the case $t > s$, things depend on the interval in which s lies, denoted by β:

$$\begin{aligned} q_2(t;\alpha|s;\beta)|_{t>s} &= \left(1 - e^{-2rs}\right) \frac{1}{s} \lambda(t) L(s;t) \int_0^s L(u;s)^2 du., \\ &= \left(1 - e^{-2rs}\right) \frac{1}{s} \lambda_\alpha L(s;t) \left(\sum_{\gamma=0}^{\beta-1} \int_{T_\gamma}^{T_{\gamma+1}} L(u;s)^2 \, du + \int_{T_\beta}^{s} L(u;s)^2 \, du \right) \end{aligned} \tag{22}$$

1.4 Integrating over time intervals

For each time interval we now have to integrate t through $[T_a; T_{a+1}]$. First the equilibrium probability:

$$\begin{aligned} q_0(\alpha) &= \int_{T_\alpha}^{T_{\alpha+1}} \lambda_\alpha L(0;t) dt \\ &= \int_{T_\alpha}^{T_{\alpha+1}} \lambda_\alpha L\left(0;T_\alpha\right) e^{-(t-T_\alpha)\lambda_\alpha} dt \\ &= L\left(0;T_\alpha\right) \left(1 - e^{-\Delta_\alpha \lambda_\alpha}\right) \end{aligned} \tag{23}$$

Next, we compute the expected time in interval β:

$$\langle t_\beta \rangle = \frac{1}{q_0(\beta)} \int_{T_\beta}^{T_{\beta+1}} t q_0(t;\beta) dt = \frac{1}{L\left(0;T_\beta\right)\left(1 - e^{-\Delta_\beta \lambda_\beta}\right)} \int_{T_\beta}^{T_{\beta+1}} t \lambda_\beta L(0;t) dt \tag{24}$$

This expression for $\langle t_\beta \rangle$ has a numerical instability for $\lambda_\beta \lesssim 10^{-3}$. We set the following asymptotic values:

$$\langle t_\beta \rangle = \begin{cases} \left(T_\beta + T_{\beta+1}\right)/2 & \text{for } \lambda_\beta < 10^{-3} \text{ and } T_{\beta+1} < \infty \\ T_\beta + {\lambda_\beta}^{-1} & \text{for } \lambda_\beta < 10^{-3} \text{ and } T_{\beta+1} = \infty \end{cases} \tag{25}$$

We can now write down equations for the off-diagonal elements of the transition

matrix, i.e. elements with $\alpha \neq \beta$. First the case $\alpha < \beta$:

$$\begin{aligned}
q_2(\alpha|\beta)|_{\alpha<\beta} &= \int_{T_\alpha}^{T_{\alpha+1}} q_2\left(t;\alpha|\langle t_\beta\rangle;\beta\right)|_{t<s} dt \\
&= \int_{T_\alpha}^{T_{\alpha+1}} \left(1-e^{-2r\langle t_\beta\rangle}\right)\frac{1}{\langle t_\beta\rangle}\lambda_\alpha \times \\
&\quad \left(\left(\sum_{\gamma=0}^{\alpha-1} L\left(T_{\gamma+1};t\right)^2 \frac{1}{2\lambda_\gamma}\left(1-e^{-2\lambda_\gamma\Delta_\gamma}\right)\right)+\frac{1}{2\lambda_\alpha}\left(1-e^{-2\lambda_\alpha(t-T_\alpha)}\right)\right) dt \\
&= \left(1-e^{-2r\langle t_\beta\rangle}\right)\frac{1}{\langle t_\beta\rangle}\lambda_\alpha\left(\left(1-e^{-2\Delta_\alpha\lambda_\alpha}\right)\sum_{\gamma=0}^{\alpha-1}\left(\frac{1}{2\lambda_\gamma}\left(1-e^{-2\lambda_\gamma\Delta_\gamma}\right)L\left(T_{\gamma+1};T_\alpha\right)^2\right)+\right. \\
&\qquad \left.\frac{1}{2\lambda_\alpha}\left(\Delta_\alpha-\frac{1}{2\lambda_\alpha}\left(1-e^{-2\Delta_\alpha\lambda_\alpha}\right)\right)\right)
\end{aligned} \tag{26}$$

where we have used

$$\int_{T_\alpha}^{T_{\alpha+1}} e^{-2(t-T_\alpha)\lambda_\alpha}\, dt = \frac{1}{2\lambda_\alpha}\left(1-e^{-2\Delta_\alpha\lambda_\alpha}\right) \tag{27}$$

Analogously we have:

$$\begin{aligned}
q_2(\alpha|\beta)|_{\alpha>\beta} &= \int_{T_\alpha}^{T_{\alpha+1}} q_2\left(t;\alpha\langle t_\beta\rangle;\beta\right)|_{t>s} dt \\
&= \int_{T_\alpha}^{T_{\alpha+1}} \left(1-e^{-2r\langle t_\beta\rangle}\right)\frac{1}{\langle t_\beta\rangle}\lambda_\alpha L\left(\langle t_\beta\rangle;t\right)\times \\
&\quad \left(\sum_{\gamma=0}^{\beta-1}\left(L\left(T_{\gamma+1};\langle t_\beta\rangle\right)^2\frac{1}{2\lambda_\gamma}\left(1-e^{-2\lambda_\gamma\Delta_\gamma}\right)\right)+\frac{1}{2\lambda_\beta}\left(1-e^{-2\lambda_\beta(\langle t_\beta\rangle-T_\beta)}\right)\right) dt \\
&= \int_{T_\alpha}^{T_{\alpha+1}} L\left(\langle t_\beta\rangle;t\right) dt \left(1-e^{-2r\langle t_\beta\rangle}\right)\frac{1}{\langle t_\beta\rangle}\lambda_\alpha \times \\
&\quad \left(\sum_{\gamma=0}^{\beta-1}\left(L\left(T_{\gamma+1};\langle t_\beta\rangle\right)^2\frac{1}{2\lambda_\gamma}\left(1-e^{-2\lambda_\gamma\Delta_\gamma}\right)\right)+\frac{1}{2\lambda_\beta}\left(1-e^{-2\lambda_\beta(\langle t_\beta\rangle-T_\beta)}\right)\right) \\
&= L\left(\langle t_\beta\rangle;T_\alpha\right)\frac{1}{\lambda_\alpha}\left(1-e^{-\Delta_\alpha\lambda_\alpha}\right)\left(1-e^{-2r\langle t_\beta\rangle}\right)\frac{1}{\langle t_\beta\rangle}\lambda_\alpha\times \\
&\quad \left(\sum_{\gamma=0}^{\beta-1}\left(L\left(T_{\gamma+1};\langle t_\beta\rangle\right)^2\frac{1}{2\lambda_\gamma}\left(1-e^{-2\lambda_\gamma\Delta_\gamma}\right)\right)+\frac{1}{2\lambda_\beta}\left(1-e^{-2\lambda_\beta(\langle t_\beta\rangle-T_\beta)}\right)\right)
\end{aligned} \tag{28}$$

where we have used

$$\int_{T_\alpha}^{T_{\alpha+1}} L(s;t) = L\left(s;T_\alpha\right) \int_{T_\alpha}^{T_{\alpha+1}} e^{-(t-T_\alpha)\lambda_\alpha}\, dt = L\left(s;T_\alpha\right) \frac{1}{\lambda_\alpha}\left(1 - e^{-\Delta_\alpha \lambda_\alpha}\right) \tag{29}$$

The complete discrete transition matrix now reads:

$$q(\alpha|\beta) = \delta_{\alpha,\beta} q_1(\beta) + q_2(\alpha|\beta) \tag{30}$$

with

$$q_1(\beta) = 1 - \sum_{\alpha \neq \beta} q_2(\alpha|\beta) \tag{31}$$

due to the column normalization of the transition matrix.

1.5 Emission Probability

An observation at location i in the genome for a pair of haplotypes (as in a single diploid genome), O_i, can be either of $O_i = \{0, 1, 2\}$, where 0 denotes missing data in either of the two haplotypes, 1 denotes a site where both haplotypes have the same allele (i.e. a homozygous genotype in case of a single diploid genome), 1 denotes a mismatch between the alleles of the two haplotypes (i.e. a heterozygote genotype in case of a single diploid genome),

The emission probabilities for exact coalescence times are:

$$e(0|t) = 1 \tag{32}$$

$$e(1|t) = e^{-2\mu t} \tag{33}$$

$$e(2|t) = 1 - e(1|t) \tag{34}$$

For discrete time intervals, we need to integrate over the conditional probability distribution in each time interval:

$$\begin{aligned}
e(0|\alpha) &= 1 \\
e(1|\alpha) &= \frac{\int_{T_\alpha}^{T_{\alpha+1}} q_0(t) e^{-2\mu t} dt}{\int_{T_\alpha}^{T_{\alpha+1}} q_0(t)\, dt} = \frac{\int_{T_\alpha}^{T_{\alpha+1}} \lambda_\alpha L(0;t) e^{-2\mu t} dt}{L\left(0;T_\alpha\right)\left(1 - e^{-\Delta_\alpha \lambda_\alpha}\right)} \\
&= \frac{\lambda_\alpha}{\left(1 - e^{-\Delta_\alpha \lambda_\alpha}\right)} \int_{T_\alpha}^{T_{\alpha+1}} L\left(T_\alpha;t\right) e^{-2\mu t} dt \\
&= \frac{\lambda_\alpha}{\left(1 - e^{-\Delta_\alpha \lambda_\alpha}\right)} \int_{T_\alpha}^{T_{\alpha+1}} e^{-(t-T_\alpha)\lambda_\alpha} e^{-2\mu t} dt \\
&= \frac{\lambda_\alpha e^{T_\alpha \lambda_\alpha}}{\left(1 - e^{-\Delta_\alpha \lambda_\alpha}\right)} \int_{T_\alpha}^{T_{\alpha+1}} e^{-(2\mu+\lambda_\alpha)t} dt \\
&= \frac{\lambda_\alpha e^{T_\alpha \lambda_\alpha}}{\left(1 - e^{-\Delta_\alpha \lambda_\alpha}\right)} \frac{e^{-2\mu T_\alpha}}{2\mu + \lambda_\alpha} \left(1 - e^{-(2\mu+\lambda_\alpha)\Delta_\alpha}\right)
\end{aligned} \tag{35}$$

and of course we have as before:

$$e(2|\alpha) = 1 - e(1|\alpha) \tag{36}$$

There are special forms of these expressions for two cases. First, if $T_{\alpha+1} = \infty$, then we have $\Delta_\alpha = \infty$, and so the expression becomes

$$e(1|\alpha)|_{T_{\alpha+1}=\infty} = \lambda_\alpha \frac{e^{-2\mu T_\alpha}}{2\mu + \lambda_\alpha} \tag{37}$$

Second, there is again a numerical instability for $\lambda_\alpha \lesssim 10^{-3}$, in which case the expression becomes

$$e(1|\alpha)|_{\lambda_\alpha \lesssim 10^{-3}} = \frac{1}{2\Delta_\alpha \mu} e^{-2\mu T_\alpha} \left(1 - e^{-(2\mu+\lambda_\alpha)\Delta_\alpha}\right) \tag{38}$$

1.6 MSMC2 Hidden Markov Model

We can now define a Hidden Markov Model (see [1] for background reading), based on PSMC' using the above defined transition and emission probabilities. For a given sequence of length L, we define the observations as $O_1 \dots O_L$. We define a forward variable $f_1(\alpha) \dots f_L(\alpha)$ by the recursion relation:

$$f_1(\alpha) = q_0(\alpha) e(O_1|\alpha) \tag{39}$$

$$f_n(\alpha) = e(O_n|\alpha) \sum_\beta q(\alpha|\beta) f_{n-1}(\beta) \quad \text{for } n = 2 \dots L \tag{40}$$

Analogously, a "backwards"-vector $b_1(\alpha) \dots b_L(\alpha)$ is defined as:

$$b_L(\alpha) = 1 \tag{41}$$

$$b_n(\beta) = \sum_\alpha e(O_{n+1}|\alpha) q(\alpha|\beta) b_{n+1}(\alpha) \quad \text{for } n = (L-1) \dots 1 \tag{42}$$

In practice, we can speed these algorithms up substantially by precomputing powers of emission-transition matrices in order to quickly skip over long regions with missing or homozygous data. This is described in [7].

We now recursively run these two variables over all chromosomes and all pairs of haplotypes. This makes it different from MSMC, which consisted of one HMM across all haplotypes simultaneously. Here we run separately over all combinations of pairs. So for example, with two diploid phased human genomes from a single population, we would run the forward-backward algorithm independently (and possibly in parallel) over 132 chromosomal pairs of haplotypes: 6 pairs of haplotypes ((1,2), (1,3), (1,4), (2,3), (2,4), (3,4)) on 22 chromosomes each.

In order to estimate parameters of our HMM (i.e. the piecewise constant coalescence rates λ_α and the recombination rate r), we use the Baum-Welch algorithm, similarly to MSMC.

We first define an objective function

$$F(\theta, \bar{\theta}) = \sum_{\alpha,\beta} \log(q(\alpha|\beta;\bar{\theta}))\Xi(\alpha|\beta, O_n, \theta) + \sum_{O',\alpha} \log(e(O'|\alpha;\bar{\theta}))\Gamma(O', \alpha; O_n, \theta) \tag{43}$$

with

$$\Xi(\alpha|\beta, O_n, \theta) = \sum_n f_n(\beta)q(\alpha|\beta)e(O_{n+1}|\alpha)b_{n+1}(\alpha), \tag{44}$$

and

$$\Gamma(O'|\alpha, \theta) = \sum_n f_n(\alpha)b_n(\alpha)e(O_n|\alpha)I(O_n = O') \tag{45}$$

where O_n denotes the entire collection of observed data across all chromosomes and analysed haplotype pairs from all individuals, θ denotes the set of parameters used in this iteration of the algorithm, $\bar{\theta}$ denotes free parameters to be varied in the maximization step of the algorithm (see below). The first term in equation 43 sums up the evidence from the observed transitions along the data, and the second sums up the evidence from the observed emissions. Both evidence matrices depend on the data and on the current set of parameters θ. Matrix Ξ is a square-matrix with as many rows and columns as there are hidden states. Matrix Γ has as many rows as there are different symbols in the alphabet (here 3), and as many columns as there are hidden states.

The fact that all haplotypes pairs from all analysed individuals and chromosomes are summed up into one objective function corresponds to a composite-likelihood across all individuals. We essentially ignore correlations of hidden states across different pairs of haplotypes, which affects the likelihood itself, but turns out in practice to yield unbiased parameter estimates.

The sum runs in principle over all sites. In practice, we sparsen this sum by selecting an equally spaced set of sites. By default, the distance between each counted site is 1000, but this can be controlled via the parameter `--hmmStrideWidth`.

The maximization step of the Baum-Welch algorithm then re-estimates the parameters by maximizing the objective function:

$$\hat{\theta} = \arg\max_{\bar{\theta}} F(\theta, \bar{\theta}) \tag{46}$$

The Baum-Welch algorithm consists of iterations of i) the forward-backward algorithm to compute the objective function, and ii) a maximization step to estimate new parameters. In the next iteration, the forward-backward algorithm is then run with the new parameters, and so forth.

After about 20 iterations, we find that the likelihood plateaus for most MSMC runs.

Note that due to the sparsening using `--hmmStrideWidth` as explained above, it can principle happen that the likelihood does not anymore strictly increase from iteration to iteration. If that is observed, we recommend to decrease the stride width. But in practice we never observe this within 20 iterations.

1.7 Combining within- and cross-coalescence rates estimates

While MSMC can estimate three coalescence rate functions simultaneously when run over genomes from two populations, MSMC2 runs over pairs of populations separately. Each run then uses a slightly different time scaling (due to different heterozygosity, i.e. allele mismatch, estimates within and across populations). For MSMC-IM, we however need three estimates of coalescence rates defined along the same time intervals.

We supply a simple python script, called `combineCrossCoal.py`, which reads in three result files from MSMC2, each from one pair of populations, and uses interpolation of the resulting piecewise constant coalescence rate estimates to merge these datasets. Details about this can be found in the accompanying README of the `msmc-tools` repository on `github.com/stschiff/msmc-tools`.

1.8 Appendix: Normalizations

In the following derivations, we define $L(t)$ to be an antiderivative of $\lambda(t)$, i.e. $L'(t) = \lambda(t)$. We will also make use of the substitution rule

$$\int_a^b g'(x) f(g(x)) dx = \int_{g(a)}^{g(b)} f(z)\, dz. \tag{47}$$

1.8.1 PSMC conditional transition probability

We have

$$q(t|s,u,m) = \lambda(t) \exp\left(- \int_u^t \lambda(v)\, dv \right) \Theta(t-u). \tag{48}$$

We need to show that the PSMC conditional probability is normalized:

$$\begin{aligned}
\int_0^\infty q(t|s,u,m)\,dt &= \int_0^\infty \lambda(t)\exp\left(-\int_u^t \lambda(\nu)\,d\nu\right)\Theta(t-u)dt \\
&= \int_u^\infty \lambda(t)\exp\left(-\int_u^t \lambda(\nu)\,d\nu\right) \\
&= \int_u^\infty \lambda(t)\exp\left(-\int_u^t \lambda(\nu)\,dv\right) \\
&= \int_u^\infty L'(t)e^{-(L(t)-L(u)} = e^{L(u)}\int_{L(u)}^{L(\infty)} e^{-z}\,dz = e^{L(u)}\left(e^{-L(u)}-e^{-L(\infty)}\right) \\
&= 1-\exp\left(-\int_u^\infty \lambda(\nu)\,d\nu\right) \\
&= 1 \quad \square
\end{aligned} \tag{49}$$

1.8.2 PSMC' conditional transition probability

We have

$$\begin{aligned}
q(t|s,u,m) = \delta(t-s)\frac{1}{2}\left(1-\exp\left(-2\int_u^t \lambda(\nu)\,d\nu\right)\right) + \\
\begin{cases}
\lambda(t)\exp\left(-\int_u^t 2\lambda(\nu)dv\right)\Theta(t-u) & \text{for } t \le s \\
\lambda(t)\exp\left(-\int_u^s 2\lambda(\nu)d\nu - \int_s^t \lambda(\nu)\,d\nu\right) & \text{for } t > s.
\end{cases}
\end{aligned} \tag{50}$$

We again need to compute the integral $\int_0^\infty q(t|s,u,m)\,dt$. We divide the integral

into three parts:

$$
\begin{aligned}
\int_0^\infty q(t|s,u,m)\,dt = &\int_0^\infty \delta(t-s)\frac{1}{2}\left(1-\exp\left(-2\int_u^t \lambda(v)\,dv\right)\right)dt+\\
&\int_u^s \lambda(t)\exp\left(-\int_u^t 2\lambda(v)dv\right)dt+\\
&\int_s^\infty \lambda(t)\exp\left(-\int_u^s 2\lambda(v)dv-\int_s^t \lambda(v)\,dv\right)dt\\
=&\frac{1}{2}\left(1-e^{-2(L(s)-L(u))}\right)+\int_u^s L'(t)e^{-2(L(t)-L(u))}dt+\\
&e^{-2(L(s)-L(u))}\int_s^\infty L'(t)e^{-(L(t)-L(s))}dt\\
=&\frac{1}{2}\left(1-e^{-2(L(s)-L(u))}\right)+e^{2L(u)}\int_{L(u)}^{L(s)} e^{-2z}\,dz+\\
&e^{-2(L(s)-L(u))}e^{L(s)}\int_{L(s)}^{L(\infty)} e^{-z}\,dz\\
=&\frac{1}{2}\left(1-e^{-2(L(s)-L(u))}\right)+e^{2L(u)}\frac{1}{2}\left(e^{-2L(u)}-e^{-2L(s)}\right)+\\
&e^{-2(L(s)-L(u))}e^{L(s)}\left(e^{-L(s)}-e^{-L(\infty)}\right)\\
=&\frac{1}{2}\left(1-e^{-2(L(s)-L(u))}\right)+\frac{1}{2}\left(1-e^{2(L(u)-L(s))}\right)+e^{-2(L(s)-L(u))}\left(1-e^{L(s)-L(\infty)}\right)\\
=&1-e^{-2(L(s)-L(u))}+e^{-2(L(s)-L(u))}\\
=&1\quad\square
\end{aligned}
\tag{51}
$$

2 MSMC-IM model

2.1 Continuous IM model

Our model is based on Hobolth et al. 2011 [2], which demonstrates that the time to the most recent common ancestor (tMRCA) of two lineages sampled from a pair of populations can be exactly computed from a matrix exponential. Hobolth et al. 2011 [2] formulate the IM model as a continuous time Markov chain.

Here we build on that work and define a two-island model by time-dependent population sizes $N_1(t)$ and $N_2(t)$ and a time-dependent continuous symmetric migration rate $m(t)$ between the two populations, discarding the clean split concept in Hobolth et al. but describe the population separation as a continuous process.

The state space of our Markov chain matches the state space from the model in Hobolth et al. for times more recent than the split time. There are five possible states of uncoalesced and coalesced lineages: S_{11} denotes two uncoalesced lineages residing in population 1; S_{12} denotes the state where one lineage resides in population 1 and the other in population 2; S_{22} denotes both lineages residing in population 2; S_1 describes the state where the two lineages have coalesced, and the single remaining lineage resides in population 1; S_2 similarly, where the single remaining lineage resides in population 2.

The state of the two lineages composes a series of states in a Markov chain. At time $t = 0$ (the present-day generation), the state of two randomly sampled uncoalesced lineages starts from either of the following three states S_{11}, S_{12}, S_{22}, and at any later time end up in any of five states S_{11}, S_{12}, S_{22}, S_1 or S_2.

We describe this evolution of the state space via a probability vector $x_n(t)$ denoting the state probability to be in state n at time t, with time counting backwards in time. We summarise that vector in bold font as $\mathbf{x}(t)$.

We summarise the transition rate between states by a matrix $\mathbf{Q}(t)$, where rows indicating the state at some time t, and columns the state one generation later. Then the matrix $\mathbf{Q}$ can be expressed in terms of a symmetric migration rate and effective population sizes (very similar to [2]):

$$\mathbf{Q} = \begin{array}{c} \\ S_{11} \\ S_{12} \\ S_{22} \\ S_1 \\ S_2 \end{array} \begin{array}{c} \begin{array}{ccccc} S_{11} & S_{12} & S_{22} & S_1 & S_2 \end{array} \\ \begin{pmatrix} \cdot & 2m(t) & 0 & \frac{1}{2N_1(t)} & 0 \\ m(t) & \cdot & m(t) & 0 & 0 \\ 0 & 2m(t) & \cdot & 0 & \frac{1}{2N_2(t)} \\ 0 & 0 & 0 & \cdot & m(t) \\ 0 & 0 & 0 & m(t) & \cdot \end{pmatrix} \end{array}$$

where $N_1(t)$, $N_2(t)$ and $m(t)$ are all time-dependent. Diagonal elements are set such that rows sum up to zero. The state probability vector in the next generation is then the product of $\mathbf{x}(t)$ and $\mathbf{Q}$:

$$\mathbf{x}(t+1) = \mathbf{x}(t) \cdot (\mathbf{1} + \mathbf{Q}), \tag{52}$$

where $\mathbf{1}$ is a diagonal unit matrix. For n generations, we get

$$\mathbf{x}(t+n) = \mathbf{x}(t) \cdot (\mathbf{1} + \mathbf{Q})^n. \tag{53}$$

We now switch to continuous time, and note that for a small time interval Δt we can write:

$$\mathbf{x}(t_0 + \Delta t) = x(t_0) \cdot (1 + \Delta t \mathbf{Q}) \tag{54}$$

Longer time segments t can then be divided into n small time intervals, and we assume Q is constant in each interval and independent from matrices in other

intervals.

$$\mathbf{x}(t_0 + t) = \mathbf{x}(t_0) \cdot \left(\mathbf{1} + \frac{t}{n} \mathbf{Q} \right)^n \tag{55}$$

In the limit of $n \to \infty$, the equation above becomes a matrix exponential:

$$\mathbf{x}(t_0 + t) = \mathbf{x}(t_0) \cdot e^{\mathbf{Q}t} \tag{56}$$

When $t_0 = 0$, we then have:

$$\mathbf{x}(t) = \mathbf{x}(0) \cdot e^{\mathbf{Q}t} \tag{57}$$

We can use this general state propagation equation to compute the conditional probability of ending up in a specific final state s_f after time t given a specific starting state s_0. For example, the probability to end in state $s_f = S_{11}$ when starting in state $s_0 = S_{12}$ would be:

$$G(s_f = S_{11}, t | s_0 = S_{12}) = \left[\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \cdot e^{\mathbf{Q}t} \right]_{S_{11}} \tag{58}$$

where we have followed the convention introduced above that the order of states in vector notation is S_{11}, S_{12}, S_{22}, S_1, S_2.

We can now use this to write down the probability of a coalescence event of the two lineages at time t, starting in on of the starting states $s_0 \in \{S_{11}, S_{12}, S_{22}\}$:

$$\mathbf{P}^{\mathrm{IM}}(t | s_0, N_1, N_2, m) = G(S_{11}, t | s_0) \cdot 1/2N_1 + G(S_{22}, t | s_0) \cdot 1/2N_2 \tag{59}$$

because in order for a coalescence event to occur exactly at time t, we require that i) no coalescence has occurred before (so we exclude final states S_1 and S_2), ii) both lineages are in the same population (so we exclude S_{12}).

2.2 Comparing with MSMC outputs

MSMC (here as a term used independently from a specific implementation like MSMC or MSMC2) estimates time-dependent effective coalescent rates λ_{ij} between a pair of lineages i and j. From these rates, we can compute the probability density for coalescence events:

$$\mathbf{P}^{\mathrm{MSMC}}(t | s_0 = S_{ij}) = \lambda_{ij}(t) \cdot e^{-\int_0^t \lambda_{ij}(t')dt'} \tag{60}$$

The basic idea behind MSMC-IM is to fit the model from equation 59 to the observed distribution from equation 60 to estimate parameters $N_1(t)$, $N_2(t)$ and $m(t)$.

2.3 Model Fitting

So far we haven't specified the form of the time-dependent parameters $N_1(t)$, $N_2(t)$ and $m(t)$. Since MSMC uses piecewise constant functions for the coalescence rates, we decided to use exactly the same method in MSMC-IM, and impose a piece-wise constant structure on our model parameters with the same time patterning as in MSMC.

We denote the time boundaries by t_i, with $i = 0 \dots n_T$, where n_T is the number of time segments, and $t_0 = 0$ is the left-most time-boundary, and $t_{n_T} = \infty$ is the rightmost time segment. Note that in practice we set $t_{n_T} = 4t_{n_T-1}$. We can then define the following χ^2-statistic across all time-segments to measure the fit deviation between the coalescent distributions from MSMC and the IM model:

$$\tilde{\chi}^2 = \sum_{i=0}^{n_T} \sum_{x_0 \in \{S_{11}, S_{12}, S_{22}\}} \frac{(\mathbf{P}^{IM}(t_i|s_0) - \mathbf{P}^{MSMC}(t_i|s_0))^2}{\mathbf{P}^{MSMC}(t_i|s_0)} \tag{61}$$

For brevity we omit the dependency on model parameters $N_1(t)$, $N_2(t)$ and $m(t)$ here. Minimization of this χ^2-statistic is numerically implemented via Powell's method (using the function `minimize(method='Powell')` from the scipy-package in python (www.scipy.org)).

2.3.1 Regularisation

We need to estimate N_1, N_2 and m for each time interval, which for the default MSMC time patterning means 96 parameters in total. It turns out that this model is overspecified for times at which the two populations have almost completely merged (as for example reflected by $M(t)$ approaching 1, see main text). To avoid over-fitting, we add two regularisation terms to the above χ^2-statistic:

$$\tilde{\chi}^2 = \sum_{i=1}^{n_T} \sum_{s_0 \in \{S_{11}, S_{12}, S_{22}\}} \frac{(\mathbf{P}^{IM}(t_i|s_0) - \mathbf{P}^{MSMC}(t_i|s_0))^2}{\mathbf{P}^{MSMC}(t_i|s_0)} + \beta_1 \int_0^\infty m(t)\mathrm{d}t + \beta_2 \sum_{i=0}^{n_T} \left(\frac{N_1(t_i) - N_2(t_i)}{N_1(t_i) + N_2(t_i)} \right)^2 \tag{62}$$

The regularization terms β_1 and β_2 are tunable, and in practice we set β_1 to 1e-8 and $beta_2$ to 1e-6 by default. This $beta_1$ value was chosen to be low enough to not affect migration rate estimates but avoid over-estimation, and the $beta_2$ value was chosen to be low enough to not affect population size estimates at time substantially before the split time, but strong enough to "pull together" the two population sizes for times very deep in the past, where all lineages have effectively merged into one population.

2.3.2 Hazard function for estimating coalescence rates from IM model

While the primary variable to use for comparison between model and data is the probability density function of pairwise coalescence times (eqs. 60 and 59), we can also compute the Hazard function from the model, to be directly compared to the pairwise coalescence rates output by MSMC: as following equation:

$$\lambda_{ij}^{IM}(t) = \frac{\mathbf{P}(t|s_0 = S_{ij}, N_1, N_2, m)}{1 - \int_0^t \mathbf{P}(t|s_0 = S_{ij}, N_1, N_2, m)} \tag{63}$$

This expression becomes numerically unstable for very ancient times, for which the denominator becomes too small.

2.3.3 Internal auto-Correction and parameter constraints

In some cases, MSMC coalescence rate estimates in the most ancient few time intervals are noisy, which can affect migration rate estimates in these windows and lead to artifacts. We therefore implemented an automatic check of the rate estimates in the most ancient time intervals before fitting with MSMC-IM, and auto-correct these values. Specifically, we check in all time segments that correspond to the last two free parameters (with the default patterning of 1*2+25*1+1*2+1*3, as in MSMC2, the last five time intervals would be checked). In these intervals, since we do not genuinely expect estimates to fluctuate much at this end of the analysis time window, we require estimates to fall within a range of $[a/1.5, a \times 1.5]$, where a is the value of the third-last free parameter in MSMC, so the time segment just before the segments that are checked. If this condition is not fulfilled, we correct the estimates in the checked time intervals to a. This autocorrection is independently performed for each pair of haplotypes analysed (so for example we independently check λ_{11}, λ_{12} and λ_{22} independently).

We also constrain parameters $N_1(t)$ and $N_2(t)$ to be below 10^7 and migration rates to be below 100, to avoid overflow issues during the fit. Furthermore, in MSMC-IM's automatic output report, we do not report estimated migration rates for times more ancient than after $M(t)$ has reached 0.999, because of the very little data that is left to infer migration rates when all but 0.1% of lineages have effectively already merged in one ancestral population.

2.3.4 Interpreting Population size estimates

In MSMC-IM, we have two populations that never merge into one ancestral population. Instead, continuous migration is used to model movement of lineages across population boundaries, and hence also coalescence events between lineages sampled across populations.

The degree to which lineages get mixed, looking back in time, can be quantified by the cumulative migration density, as defined in Methods as

$$M(t) = 1 - e^{-\int_0^t m(t')\mathrm{d}t'} \tag{64}$$

In recent times, where $M(t) \ll 1$, population sizes parameters $N_1(t)$ and $N_2(t)$ correspond closely to the inverse coalescence rates $1/\lambda_{11}(t)$ and $1/\lambda_{11}(t)$ estimated by MSMC. However, as $M(t)$ approaches 1, the interpretation of these parameters differs from what one would normally call an "ancestral population size" in a clean-split model: In our model, we maintain two separate populations, so that with probability 1/2, two lineages will be in separate populations and cannot coalesce. Therefore, the effective coalescence rates in MSMC-IM for times at which $M(t) \to 1$, is half the rate expected for an ancestral population with size $N_1(t)$ or $N_2(t)$.

Therefore, for $M(t) \to 1$, a meaningful estimate for the effective "ancestral" population size would be $2N_1(t) \approx 2N_2(t)$. We therefore found it useful to report "corrected" population size estimates defined as

$$\begin{aligned} \mathbf{N_1'}(\mathbf{t}) &= (1 - M(t))N_1(t) + M(t)2N_1(t) \\ \mathbf{N_2'}(\mathbf{t}) &= (1 - M(t))N_2(t) + M(t)2N_2(t) \end{aligned} \tag{65}$$

References

[1] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[2] Asger Hobolth, Lars Nørvang Andersen, and Thomas Mailund. On computing the coalescence time density in an isolation-with-migration model with few samples. *Genetics*, 187(4):1241–1243, April 2011.

[3] Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, July 2011.

[4] Anna-Sapfo Malaspinas, Michael C Westaway, Craig Muller, Vitor C Sousa, Oscar Lao, Isabel Alves, Anders Bergström, Georgios Athanasiadis, Jade Y Cheng, Jacob E Crawford, Tim H Heupink, Enrico Macholdt, Stephan Peischl, Simon Rasmussen, Stephan Schiffels, Sankar Subramanian, Joanne L Wright, Anders Albrechtsen, Chiara Barbieri, Isabelle Dupanloup, Anders Eriksson, Ashot Margaryan, Ida Moltke, Irina Pugach, Thorfinn S Korneliussen, Ivan P Levkivskyi, J Víctor Moreno-Mayar, Shengyu Ni, Fernando Racimo, Martin Sikora, Yali Xue, Farhang A Aghakhanian, Nicolas Brucato, Søren Brunak, Paula F Campos, Warren Clark, Sturla Ellingvåg, Gudjugudju Fourmile, Pascale Gerbault, Darren Injie, George Koki, Matthew Leavesley, Betty Logan, Aubrey Lynch, Elizabeth A Matisoo-Smith, Peter J

McAllister, Alexander J Mentzer, Mait Metspalu, Andrea B Migliano, Les Murgha, Maude E Phipps, William Pomat, Doc Reynolds, François-Xavier Ricaut, Peter Siba, Mark G Thomas, Thomas Wales, Colleen Ma'run Wall, Stephen J Oppenheimer, Chris Tyler-Smith, Richard Durbin, Joe Dortch, Andrea Manica, Mikkel H Schierup, Robert A Foley, Marta Mirazon Lahr, Claire Bowern, Jeffrey D Wall, Thomas Mailund, Mark Stoneking, Rasmus Nielsen, Manjinder S Sandhu, Laurent Excoffier, David M Lambert, and Eske Willerslev. A genomic history of aboriginal australia. *Nature*, September 2016.

[5] Paul Marjoram and Jeff D Wall. Fast "coalescent" simulation. *BMC Genet.*, 7:16, January 2006.

[6] Gilean A T McVean and Niall J Cardin. Approximating the coalescent with recombination. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 360(1459):1387–1393, July 2005.

[7] Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.*, 46(8):919–925, August 2014.

14.2. Supplementary Materials of paper B

Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/24/eaaz0183/DC1

Supplementary Materials for

Ancient genomes reveal complex patterns of population movement, interaction, and replacement in sub-Saharan Africa

Ke Wang, Steven Goldstein, Madeleine Bleasdale, Bernard Clist, Koen Bostoen, Paul Bakwa-Lufu, Laura T. Buck, Alison Crowther, Alioune Dème, Roderick J. McIntosh, Julio Mercader, Christine Ogola, Robert C. Power, Elizabeth Sawchuk, Peter Robertshaw, Edwin N. Wilmsen, Michael Petraglia, Emmanuel Ndiema, Fredrick K. Manthi, Johannes Krause, Patrick Roberts, Nicole Boivin*, Stephan Schiffels*

*Corresponding author. Email: boivin@shh.mpg.de (N.B.); schiffels@shh.mpg.de (S.S.)

Published 12 June 2020, *Sci. Adv.* **6**, eaaz0183 (2020)
DOI: 10.1126/sciadv.aaz0183

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/6/24/eaaz0183/DC1)

Supplementary Materials

Supplementary Text

Text S1. Ethical statement

This is an ethical statement regarding ancient individuals from Kindoki and Ngongo Mbata site in the Democratic Republic of the Congo (DRC). The KongoKing research project was conducted in the DRC authorized by the "Arrêté ministériel n°0115/CAB/MIN/JSCA/2012" dated August 8, 2012. Each year a separate document was signed between the KongoKing project mission Director and the Minister for Culture to authorize exportation for scientific study of artifacts and ecofacts. Furthermore, at the local village level, all surveys and excavations had to be accepted by the local community, often paying out a "droit de passage" to the land owner(s). Each time an ancient burial was encountered during the fieldwork, further work could be carried out only by organizing with the local specialized man the ritual to appease the dead spirit(s). At the end of any burial excavation, before filling up, another ceremony was conducted to thank the cooperation of the deceased person leading to not having had any incident.

Text S2. Archaeological information on newly reported individuals

Here we describe the archaeological sites at which our ancient individuals were found.

Lukenya Hill, GvJm 202, Kenya

Lukenya Hill is an Archaean Basement gneiss inselberg located in southern Kenya (−1.465°, 37.067°), in the Athi-Kapiti Plains east of the Central Rift Valley and just outside the current extent of Nairobi. The Lukenya inselberg has several dozen excavated archaeological sites including rockshelters with Middle and Later Stone Age components (*69*) as well as open-air Pastoral Neolithic sites (*70*, *71*) as well as rock-art occurrences and remains of historic to recent Maasai meat-feasting rituals.

Several major Neolithic sites are located at elevations of between 1600-1700m on Lukenya Hill, including GvJm44 which yielded charcoal radiocarbon dates of 3290±145, (GX5348) (*71*). Along with finds of "Nderit" pottery styles better known from 5000-4000 BP in Lake Turkana of northern Kenya, Lukenya Hill has been considered a major loci for the earliest occurrences of the Pastoral Neolithic in southern Kenya (*25*). An additional major Pastoral Neolithic (PN) site is GvJm184, which yielded slightly later dates of 2716-1735 cal BP (*71*, *72*). GvJm184 has yielded typical Savanna Pastoral Neolithic (SPN) pottery styles as well as remains of domesticated cattle, sheep and goat.

The site of GvJm202 is a rockshelter site including several human burials located nearby to the PN site of GvJm184, and so the burials were believed to be attributed to the Pastoral

Neolithic phase (*72*). Excavations found remains of at least 6 individuals consisting of 5 adults and 1 sub-adult (*73*, *74*). A sample from Skeleton C (labeled in original excavations) and two samples of loose human remains from Context 11 and Context 1 were sampled for aDNA. Samples from Skeleton C and the loose remains yielded aDNA sequences.

The loose remains were labeled in this study "LUK001" and the Skeleton C sample labeled in this study "LUK003" with dates of 3610–3460 cal BP and 3635–3475 cal BP., respectively. Given that these were very temporally close date ranges with similar genetic composition, both individuals are given the population label "Kenya_LukenyaHill_3500BP". LUK001 was found to be a male and LUK003 was found to be female (*contra* osteological indicators that were ambiguous-to-slightly-male, see (*73*, *74*)). With the exception of two individuals from PretteJohns Gully individual dated to ~4200 BP and presumed to be associated with an early movement of herders (*4*), the Lukenya individuals here represent the oldest examples of the migration event(s) responsible for the more pronounced expansion of herding detected all across Kenya and Tanzania (*3*, *4*).

Hyrax Hill, GrJj25, Kenya

Hyrax Hill is a major multi-component archaeological site and National Monument on the northeast shore of Lake Nakuru, Central Rift Valley, southern Kenya. Hyrax Hill has three major locales, of which two (Hyrax Hill I and Hyrax Hill II) have seen major excavations. Based on archaeological materials and pottery styles, primary deposits at the site belong to the Savanna Pastoral Neolithic (SPN) and Iron Age phases (*75*). Excavations by Mary Leakey from 1937-1938 revealed Neolithic and Iron Age village deposits which included hut and pit structures. The Neolithic component of the site included one of the largest formal cemeteries known for the Late Holocene of southern Kenya, containing over 18 burials within a central mound. Burials were shallow but otherwise variable, sometimes containing secondary burials and sometimes with multiple individuals apparently in a single pit. Large flat stones were placed over burial pits with stone circles resembling very small cairn-type features (*75*). According to the original report, nine individuals thought to be women were buried with large grinding stones or "platters".

Capping burials with shallow mounds and large stone slabs, and the heterogeneous style of burials, bear close similarities to the earlier Pastoral Neolithic traditions documented from 5000-4000 BP around Lake Turkana (*76*). Recovery of sedge beads and pestles, and obsidian blades associated with the burials is also comparable to the burial goods at the mortuary site of Njoro River Cave dated to c. 3100 BP around 20km west of Hyrax Hill. These two packages of mortuary characteristics are associated with the Nderit and Elmenteitan pastoralist traditions respectively, showing considerable overlap and interrelation between these archaeological entities that is now known from genetic results (*4*). SPN contexts at Hyrax Hill had not previously been radiocarbon dated, and common speculation was that site dated to around 2000 BP (*73*, *74*). A second burial area at Hyrax Hill consists of a double burial in a stone cairn associated with pit features thought to be Iron Age cattle enclosures. Additional excavations have taken place at the site, but have not yet been published.

The individual samples for this aDNA study all come from the SPN burial grounds excavated by Mary Leakey. The only individual to yield high aDNA coverage was found to be male and

was directly dated to 2365-2305 calBP, confirming estimates for the age of the site and its cultural attribution as Pastoral Neolithic.

Molo Cave, GoJi3, Kenya

Molo Cave is an archaeological occurrence in the Mau Escarpment around 50km west of Lake Nakuru in the Central Rift Valley of southern Kenya. The remains of three individuals were salvaged by Mary D. Leakey and deposited in the National Museums of Kenya, but there is very little archaeological or accession data to speculate on its cultural affiliation or context (*77*). It has been believed to likely date to the Pastoral Neolithic period (*73*). aDNA was recovered from two of the three individuals from this site, yielding dates of 1415-1320 calBP (MOL001) and 2110-1990 calBP (MOL003), confirming a likely Pastoral Neolithic attribution. The genetic composition of MOL003 strongly associates it with other Pastoral Neolithic samples, and MOL001 is somewhat intermediate between PN and Pastoral Iron Age samples. Coupled with its late date, it is possible MOL001 reflects admixture between PN and PIA populations at this later time.

Nyarindi Rockshelter, GqJc13, Kenya

Nyarindi is a rockshelter site along the southeast of Lake Victoria, Nyanza, Kenya. The site was informally excavated between 1939 and 1941 by a British deacon (surname Owen) (*78*). Deacon Owen did appear to record stratigraphic levels to some degree, but reported only on the early “Smithfield” levels at the lowest part of the archaeological sequence, which seem to be an Early Stone Age component. The upper layers are poorly recorded, however Owen recovered the remains of possibly five individuals, mostly in the form of mandibles and cranial fragments, with only a few post cranial fragments. The only artifacts from the upper portions of the excavations are a few pottery sherds and quartz stone tools. The quartz tools are non-descript but likely derive from Later Stone Age industries, and the pottery is typical Late Iron Age roulette decorated styles.

Two of the individuals (labeled here NYA002 and NYA003) of the three sampled by this study yielded genomic data. These were found to be a female and male respectively. A direct date on NYA002 came out to 3555-3375 calBP. Given this date and the genetic affiliation with other eastern African hunter-gatherer samples in this study and previously reported, it is most likely this individual is related to the Kansyore fisher-forager traditions of Lake Victoria. Pottery collected from the site must therefore post-date the burials by over 2000 years. As is expected, these individuals cluster with other hunter-gatherer samples from around Lake Victoria, including the Kenya_Kakapel_3900BP sample (*4*).

Kakapel, Kenya

The Kakapel (also called Kakapeli) Rockshelter site is a granitic tor within the Chelelemuk Hills south of Mount Elgon in North Teso, Busia County, of western Kenya. Iron Age rock art is visible within an overhang that is approximately 5m deep and 4-10m tall along the southern edge of the pluton. The area surrounding the rock art has extensive evidence for prehistoric occupation, and is where archaeological excavations have been focused. Kakapel is currently a protected site managed by the Trust for African Rock-Art (TARA) and the National Museums of Kenya and is surrounded by mixed-open forest and small patches

of pasture where cattle are occasionally left to graze. The entirety of the region surrounding the site is under agricultural cultivation, primarily for corn.

There was no record of excavations at the Kakapel Rockshelter site (except some sub-surface testing mentioned by Dr. Odak previously) until 2012 when a National Museums of Kenya(NMK) team initiated testing with a single 1x1 trench located near the rock art panel(Trench I).In 2015, the NMK returned for more extensive excavations including opening two new trenches (Trench II and Trench III). The NMK encountered human remains in both units (at ~35cm below surface in Trench II and ~1m below surface in Trench III). The individual (Burial 2) in Trench II was well preserved with almost all elements present, whereas the individual in Trench III (Burial 1) exhibited in-situ taphonomic crushing that destroyed much of the axial skeleton. Petrous bone from the Burial 1 individual, a tooth and right first metacarpal from the Burial 2 individual and one isolated tooth from a third individual found near Burial 2 were selected for aDNA sampling.

Kenya_Kakapel_3900BP

This individual (Burial 1) was in a primary, articulated burial in a shallow pit in Trench III. The body was flexed on its left side, oriented south-southeast to north-northwest and facing west. The skeleton was highly fragmentary and incomplete. Three iridescent perforated shell fragments were recovered near the face, with 25 disc beads made from various raw materials (including ostrich and land snail) were recovered near the legs. Contexts surrounding the burial contained pottery attributed to the Kansyore fisher-forager traditions around Lake Victoria.

Skeletal morphology suggests a male individual (confirmed by aDNA analyses) who likely died between 20-30 years old. The third molars had erupted and finished development, but were lightly worn. Dental health was otherwise good.This individual was directly dated to 3974–3831 cal. BP (3584±28 bp, SUERC-86057). These dates are consistent with associated material culture suggesting a connection with Kansyore fisher-forager traditions that existed around Lake Victoria through the Holocene. Isotopic analyses of tooth enamel from the individual yielded a delta 15N/14N of 7.76 and a delta 13C/12C of -15.7, far lower than the -4 to -8 values from fauna from the same levels. While the individual appears to have had a major protein component of the diet, he was either consuming other faunas that grazed on C3 plants or had a diet largely based in direct C3 plant consumption.

Kenya_Kakapel_300BP

This well-preserved individual (Burial 2) was in a primary burial in Trench II. The body was semi-flexed on their left side, oriented southeast-northwest with the head facing south-southwest.There was no evidence of a burial pit, although the remains were placed between large pieces of roof fall.This individual was directly dated to 309-145 cal. BP. Skeletal morphology is consistent with a female individual (confirmed by aDNA) who died in middle adulthood. Complete fusion of the sphenooccipital synchondrosis and medial clavicle indicate an age over 30, while age estimates from the pubic symphyses, pelvic auricular surfaces and first rib suggest an age range of mid 30s-50s (*79*, *80*). Stature estimates based on the complete right humerus, radius, ulna, femur, and fibula range between 153.5 ± 4.25cm and 158.0 ± 5.05cm, with a mean of 156 cm or about 5’1.5 (*81*).

Kenya_Kakapel_900BP

This individual is represented by an upper right lateral incisor found in the Burial 2 fill but that did not belong to that individual. The tooth was directly dated to 910-736 cal. BP. An upper right canine, potentially from the same individual, was recovered from a greater depth. While this individual cannot be attributed to any clear material culture, the radiocarbon date places it within the middle-to-late phases of the Iron Age, probably with the Roulette pottery traditions of the African Great Lakes region for this time.

Munsa, Uganda

Munsa is one of several sites with earthworks in western Uganda first reported in the colonial period (*82*). The earthworks comprise systems of ditches, commonly up to 4 meters deep and often encircling one or more hills and dated to approximately 500 calBP. The function of the earthworks is uncertain (*83*). Those at Munsa, which surround a flat-topped rocky hill, were first mapped by Lanning (*84*). Excavations at the center of the site in the 1990s (*34*) uncovered numerous grain-storage and other pits, some with burials, as well as an iron-smelting furnace (*85*), numerous potsherds, bones of cattle and other animals, and occasional iron artifacts, glass beads, and grinding stones. Many of these features and artifacts date to the centuries immediately prior to the construction of the earthworks, which are linked to oral traditions of an ancient kingdom (*86*, *87*).

Individual H.s.#7, analysed here, was recovered from Context 250 (a pit fill) in Unit 101E/84N at Munsa A, i.e. on top of the hill at the center of the earthworks. The skeleton was found in a bell-shaped pit that was likely used originally for grain storage. The articulated skeleton is that of a woman aged around 35 – 50, and our genetic analysis confirmed the female sex. The skeleton was laid out on a NW-SE axis, extended, lying on her back with the head looking left (north). There was an iron bangle associated with the body. The pit contained another female human skeleton, with a few associated glass beads, lower down that had been disturbed when H.s.#7 was interred. Also in the pit were charcoal, sherds, grindstone fragments, a few pieces of iron slag, and animal bones, as well as an almost whole pot and large quern at the same level as H.s.#7.

Not enough material for direct dating of H.s. #7 was available to us. An AMS date on charcoal from the same context as the lower burial (#13) resulted in a weighted average of 955 +/- 40 bp. There is also an AMS date on charcoal from a context above H.s.#7, probably around the very top of the fill of the pit of 485 +/- 45 bp . Collagen from H.s. #13 yielded an AMS date of 345 +/- 45 bp, although based on low amounts of collagen. This collagen date of 345 bp is younger than that of the charcoal date from much higher in the pit fill of 485 bp, with a small overlap at the sigma-2 range the late 15th century AD. It is possible that charcoal was added to the pit fill that came from an older context elsewhere on the site. Given also dates from pits nearby, we estimate the date of H.s. #7 therefore to the 14th – 16th century AD, approximately contemporary with the construction of the earthworks.

Matangai Turu Northwest, the DRC

The rock shelter site of Matangai Turu Northwest is located at 800 m above sea level in the Ituri rainforest, the DRC (*88*). This 7 metre high granite shelter lies 95 m from the Andeilu river and was known to have been recently occupied by Efe foragers when it was excavated

in the late 1990s by Julio Mercader and colleagues (*88*). The team excavated a partial human skeleton during these excavations from Level 5, directly dated to the Late Holocene (813 ± 35 ^{14}C years BP; 1218-1277 AD calibrated age (1 sigma) (UtCNr 5074). The skeleton was associated with lithics identified as 'Late Stone Age' type, animal bone and shell remains from wild taxa, fruit endocarps from forest trees, and phytoliths from tropical forest plants. Phytolith analysis indicated that the habitat was dense tropical forest, without evidence of domesticated food(*89*). Overall, this individual was considered to be heavily reliance on foraging, while tooth health, modifications, as well as stature were used to tentatively suggest that this individual came from a population with a "pygmy" phenotype (*88*). Intriguingly, however, associated findings of Late Iron Age ceramics and an iron burial good hinted at the possibility of affiliation with Nilotic, central Sudanic farming populations (*88*, *90*).

Kindoki, the DRC

Kindoki was excavated between 2012 and 2014 under the direction of B. Clist as part of the KongoKing project (2012-2016, ERC Starting Grant n° 284126) led by K. Bostoen at Ghent University, Belgium. The site (-5.086°,15.129°) is located in the current-day Congo Central province of the DRC, 95km southwest of Kinshasa and 10km northwest of Kisantu (*91*, *92*). The excavations on the large hilltop – some 537 m^2 in total – were focused on the Late Iron Age and on historical domestic layers and pits related to the Congo kingdom. The Iron Age sequence with 16 ^{14}C dates starts with Kindoki ware whose producers settled on the hilltop from circa calAD 1300 to 1450 (*93*), i.e. before the arrival of the Portuguese at the Congo mouth on the Atlantic Coast in 1482. Slightly later, a continuous occupation extends from the late 15th century to the early 19th century. The higher number of artifacts during the 17th-19th centuries combined with the presence of a cemetery of 11 tombs suggests that Kindoki was only then the center of Mbanza Nsundi, the capital of the Kongo kingdom's Nsundi province. According to oral traditions, the Kongo kingdom would have been founded in the 13th century (*94*). In 1491, the Kongo king Nzinga a Nkuwu, was the first to convert to Christianity and to become João I of Kongo, followed a few weeks later by his son Mvemba a Nzinga, becoming Afonso. In 1495, Afonso was exiled to Mbanza Nsundi together with Portuguese missionaries. The town then turned into a centre of local Christianity. After his enthronement in 1509 as king Afonso I, he developed Christianity into a royal cult to which Kongo nobles adhered.

The Kindoki cemetery was excavated in 2012 (tombs 9 and 13) and 2013 (tombs 1-2, 4-8, 11-12) (*91*, *92*). It is interpreted as being the burial site for governors of the Nsundi province and their close relatives from the second half of the 17th century to the first half of the 19th century. The 11 tombs were constructed in close proximity to each other and oriented north-east/south-west at 220°, most of them in a similar way: a rectangular pit dug down to circa 2m, a rectangular surface demarcation of stones either lying flat or on their side, a topping pavement with more or less well-dressed stones, and a covering cairn of stones of various sizes and shapes on top of the rectangular structure. Where identifiable, the deceased were deposited on their back, in an extended position. According to the full skeletal analysis (*95*, *96*), parietal bones were only preserved in five tombs: 1, 7, 8, 9 and 11. DNA results for KIN003, KIN004 and KIN002 come respectively from tombs 9, 1 and 8.

Tomb 9, a 35-40 years old man of 1.61m estimated height (*95*, *96*), with 1 musket on his left side, 2 iron bracelets, 18 wound Venetian glass beads with floral inlays. Charcoal found on his right side was dated by Beta-333285: 190+/-30 bp, 1665-1950 calAD. Combination of artifacts and ^{14}C gives this more precise chronology: 1690-1725 (musket), 1725–1850 (beads), 1665–1817 (^{14}C; 66 % probability) (*91*, *97*–*99*). The new ^{14}C date on bone is OxA-37354: 172 ± 20 bp, 1672-1950 calAD (36% probability: 1672-1744; 40% probability: 1796-1895), confirming the former one on charcoal.

Tomb 1 contained a 30-35 years old man (*95*, *96*) buried without any associated artifact. Specific wear on several teeth points to a regular use of a smoking pipe, probably in clay. OxA-37355 is the first ^{14}C date obtained for this grave: 241 ± 20 bp, 1650–1799 calAD (67% probability: 1735-1799).

Tomb 8, an adult of *circa* 40 years old, osteologically determined as female with 1.57m estimated height (*95*, *96*) buried with 1,140 wound red-on-white glass beads, 14 wound pentagonal blue glass beads, 1 wound round white glass bead, 1 wound round blue glass bead, 3 internally silvered blown glass beads, 1 copper bead, 32 crotal bells, 660 *Pusula depauperata* sea shell beads, 1 *Tympanotonus fuscatus radula* mangrove shell, 1 iron anklet, 1 iron necklace, 1 copper chain, 1 gold chain, and large parts of a shroud (*91*, *97*, *99*). According to the glass beads types, the burial dates to 1825–1845. The new ^{14}C date is OxA-37353: 217 ± 20 bp, 1656–1805 calAD (76% probability: 1727-1805), which is considerably older.

While the molecular sex is male, several pieces of archaeological evidence are more typical for the burial of a female, first and foremost the measurements carried out on the mandible, the teeth and the long bones (*95*, *96*). Moreover, the funeral material of Kindoki graves 4, 5, 6, 7 and 12 containing swords (as well as four graves in Ngongo Mbata's church cemetery, see below) is notably different from that of Kindoki tombs 8 and 11 with hundreds of glass and shell beads, sometimes associated with thick iron anklets (see also several graves in two cemeteries at Ngongo Mbata and the Kulumbimbi church burial in Mbanza Kongo). The first type of tombs has commonly been identified as male, the second type as female in line with ethnographical studies from the 19th and early 20th century.

Ngongo Mbata, the DRC

Ngongo Mbata was excavated between 2012 and 2015 (total of 847,5 m^2) under the direction of B. Clist as part of the KongoKing project (2012-2016, ERC Starting Grant n° 284126) led by K. Bostoen at Ghent University, Belgium. The site (-05.806°, 015.124°) is located in the Kongo Central province of the DRC, 14 km north-east of Kimpangu and some 8 km from the Angolan border located to its south (*100*, *101*). The 18 ^{14}C dates obtained point towards human presence during the Late Stone Age and Late Iron Age, i.e. the hunter-gatherers around 9,000-8,000 bp and a settlement probably starting in the 16th century AD. According to the historical records available (i.e. chronicles and maps), Ngongo Mbata was the main and most affluent center of the Kongo kingdom's Mbata province in the 17th century. It developed to the second largest town in the kingdom because it was an important marketplace on a long-distance trade route connecting it to the kingdom's capital and the Atlantic coast. It was a centrally located thoroughfare in-between the ocean harbors in the west and the Kwango River valley in the east. At its apogee, Ngongo Mbata covered 50 hectares, at least. Along with Kongo people of different descents, it hosted Europeans of diverse origins (Portuguese, Spanish, German, Dutch), mostly merchants but also clergy. Ngongo Mbata is unique in that it had a stone church dated by excavations to the second quarter of the 17th century. Such stone churches were then only found in Mbanza Kongo, the Kongo kingdom's central capital in northern Angola and in the Portuguese territories south of it. In the DRC, the Ngongo Mbata church is the oldest surviving one.

Four cemeteries have been identified at Ngongo Mbata. Cemetery n°1, within the walls of the early 17th century stone church, was excavated in 1938 (*102*) and 2014 (*101*), and consists at least of 38 burials dating back to the 17th and 18th centuries. Cemetery n°2, located some 200 meters south-west of the church, contained at least four tombs, all very close one to the other. The only tomb studied is certainly younger than 1692 as evidenced by a 20 *reis* Portuguese coin minted between 1692 and 1699, which was found alongside three crucifixes. The tomb probably dates back to the early 18th century (*101*). Cemetery n°3, found to the south-west of the church on the southern side of a large plaza set up to the west of the church entrance, consisted of at least three tombs widely spaced, each marked by a pavement of large stones set over a tomb dug down to a depth of about 1.6m (*101*). Tomb 1 contained an adult male whose well-preserved skeleton was chosen for DNA identification. Cemetery n°4 was found while excavating large pits to the south of the church extending under a more recent house, probably a priest's residence (*101*). Three burials were identified, one set up in a refuse pit whose filling started after circa 1630 AD when the church was built.

Tomb 1, cemetery n°3, a man of circa 20 years of age and 1.7m estimated height (*95*). The rectangular pit was dug to 1.6m deep. The deceased lay on his back in an east-west position, as in all other burials studied at Ngongo Mbata, with two small glass beads near the neck (*99*, *101*). According to Christian customs of that time, the eastward orientation of his head suggests that he was a priest. After starting filling the burial, stone blocks were set up at a depth of 1,4m, i.e. on top of the body. Specific wear on several teeth points to the regular use of a smoking pipe, probably in clay. OxA-37363 is the first ^{14}C date obtained for this grave: 211 ± 21, 1657–1809 calAD (74% probability: 1724-1809).

Xaro, Botswana

Xaro is located on a former island silted to form a peninsular in the Okavango River Panhandle some 40km northwest of Nqoma. The site was excavated by Wilmsen and Denbow (1983) and subsequently by Wilmsen and Thebe. The pottery has distinctive single/multiple-row fingernail and stylus punctate motifs, multiple bands filled with incised or comb-stamped lines, as well as rhomboidal incised motifs; these motifs are identical to those found in the lower, Divuyu, levels of Nqoma securely dated to the 7th-8th centuries as well as at Ruuga some 230km upstream in Namibia firmly dated (by charcoal in sherds) to the 6th-7th centuries (*103*).

Not enough charcoal was recovered from Xaro to attempt reliable radiocarbon dating; however, seriation of Xaro pottery motifs confirms that the Divuyu-Ruuga date range also applies to Xaro. The site has been subject to periodic inundation which has destroyed all organic matter, including bones, so little can be said about subsistence. This inundation also contaminated charcoal to the extent that reliable radiocarbon dating is not possible, but the Divuyu levels at Nqoma are well dated to the 7th-8th centuries and Ruuga is securely dated by charcoal in sherds to the 6th-7th centuries. Seriation of the pottery confirms that this date range also applies to Xaro. Unlike at Nqoma, Xaro has no stone tools or other artifacts that would indicate an earlier hunter-gatherer component, and the site is considered to be related to early farmer occupations.

Two burials were excavated at Xaro located outside of the residential area of the site. Both were individual burials of adults found in flexed positions. As is typical for Iron Age burials, the individuals were interred with pottery, which matched the typical Early Iron Age styles found throughout the deposits at Xaro. The individual sampled for this analysis was an older

adult (likely 40-50 years old); the skull has morphological and craniometrical characteristics described as similar to both recent Khoisan and Bantu-speaking peoples of southern Africa.

Isotopic analyses of the Xaro individuals revealed very low $\delta^{13}C$ from samples of bone collagen (between -16.6 and -16.9) (*104*). The strong C3 values of the Xaro individuals compared to other Early Iron Age populations was interpreted as evidence for a substantial contribution of freshwater fish to the diet, which presents C3/C4 ratios similar to diets of C3 pathway plants (*104*). Given the sites location along the Okavango River, this hypothesis seems probably and the strong freshwater fish signature likely indicates that an aquatic carbon reservoir effect would impact the accuracy of any direct dates on human bone from the site.

Nqoma, Botswana

Nqoma is an open-air site on a low plateau of the Tsodilo Hills in northwestern Botswana some 50km west of the Okavango Delta. The Hills are composed of almost unlimited quantities of red specular haematite and sparkly micaceous schist; both minerals are highly valued as cosmetics for which hard rock adit mining was carried out on a massive scale, most intensively during the period AD 750-1025 (*105*). Nqoma was excavated by Denbow and Wilmsen from 1979-1980 and in 1985 with testing in several areas of the site covering roughly 100 by 200 sq. m (*106*, *107*).

Excavations yielded abundant archaeological remains, including stone tools, metal artifacts, animal bones, macrobotanical remains, Zhizo glass beads, and marine shells from the Indian Ocean trade (*106*). Much of the archaeology was concentrated in a stratified midden deposit that varied from .5 to 1 meter in thickness, below which sediment was found to be largely sterile. Radiocarbon dates on charcoal across the site revealed a consistent sequence from 1200-750 BP (*43*). Analysis of the faunal remains revealed that herding of sheep and goat was present by c. 1400 BP, with an increasing emphasis on cattle herding becoming apparent after 1150 BP (*43*). Wild fauna and large proportions of fish remains and freshwater molluscs were present throughout the sequence, indicating a strong reliance on wild resources. The riparian component may also reflect exchange with groups living along the nearby Okavango River and/or Delta. Excavations also yielded a diverse plant diet including wild marula and mongongo nuts, as well as domesticated sorghum and pearl millet (*104*, *106*).

Three human burials were excavated at Nqoma, all originating in the north-eastern section of the site. Human burials at the site include an infant interred between two parallel stone slabs, and an adjacent burial of a young woman with features that suggested affinities to Khoisan populations (*106*). This latter burial was also laid in a flexed position along the left site with the body oriented westward such that the head was directly facing the infant burial (*106*).

Below these burials was a previous burial of an older adult woman with skeletal features identified as “Bantu”, placed in an upright flexed position facing eastward (*106*). This individual from the earlier occupations of the Nqoma site was the one sampled for ancient DNA in this study. The burial included Nqoma style pottery. The burial had an intact elaborately decorated shallow ceramic bowl at her feet; this bowl was covered by two

segments from much larger vessels. A radiocarbon date of c. 980 BP was obtained from associated charcoal. While isotopic data indicated a diet including a major C4 component (*104*), the presence of fish remains throughout the Nqoma midden raises similar questions to Xaro regarding reservoir effects, and so the associated radiocarbon date for this burial pit and related material culture provide the most reliable age estimation.

Taukome, Botswana

Taukome Hill is a basalt inselberg some 700km crow-fly kilometers southeast of Nqoma in the hardveld of eastern Botswana excavated by Denbow in 1979 (*108*). Its most prominent feature is a large midden about 70m in diameter and 1.5m in depth indicating an extended occupation history initially confirmed by uncorrected radiocarbon readings and later calibrated to ca. AD 670-880 at the bottom and ca. AD 980-1150 at the top (c. 1240-995 BP).

This sequence encompasses the Zhizo (called Taukome, after this site, in Botswana) and Toutswe facies continuum of the larger Shashe-Limpopo region in which the site is a westernmost location near the Kalahari margin. Zhizo is a major division of the Nkope Branch of the Eastern Stream Urwere Tradition (*109*); thus, Taukome-Toutswe wares are markedly different from the Western Stream Kalundu wares at Nqoma indicating they were made by distinctly different Bantu-speaking peoples. A few Zhizo beads were found in the upper levels as were typically Zhizo clay human figurines. Cattle and sheep/goats were prominent in the economy, constituting 82% of the fauna based on MNI. The presence of grain bins at the site also indicate that agriculture was important (*108*).

Denbow excavated five burials including three adults and two juveniles within a kraal (livestock pen) feature at the site (*42*). The individual sampled here was an adult male, estimated to be around 40 years of age. Isotopic analyses of the Taukome burials demonstrated $\delta^{13}C$ values between -7.4 and -8.2, and $\delta^{15}N$ values between 9.5 and 10.3, largely consistent with the nearby Iron Age site of Kgaswe (*42*). Based on comparisons with isotopic signatures of fauna from these sites, it is likely that despite evidence for agriculture the isotopic signatures at Taukome can be explained through consumption of domesticated livestock (*42*).

Supplementary Figures

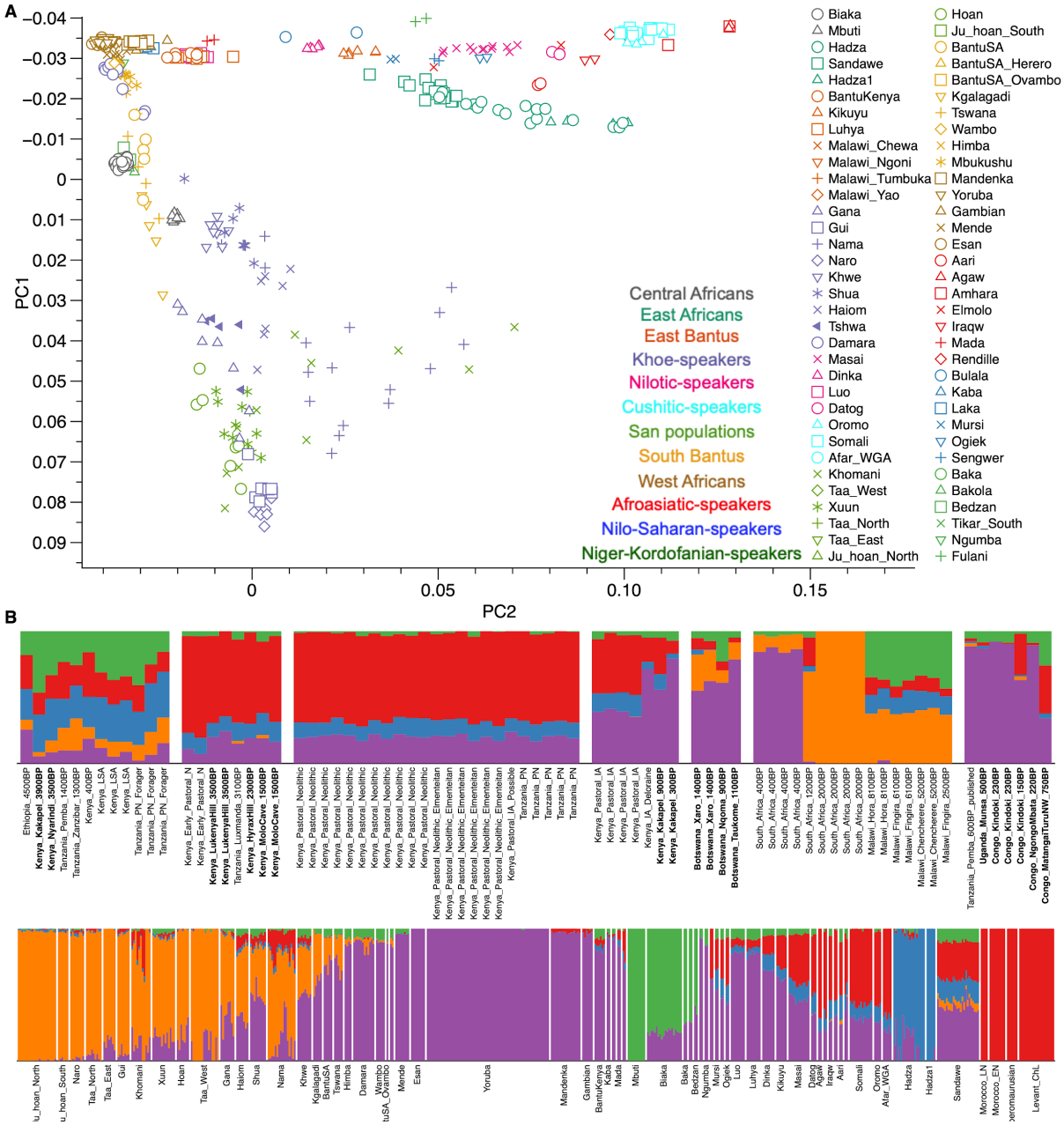


Figure S1. (A) PCA of present-day African and Levantine populations used in Fig. 2. (B) Admixture clustering analysis (K = 5) using all populations shown in PCA in Fig.2.

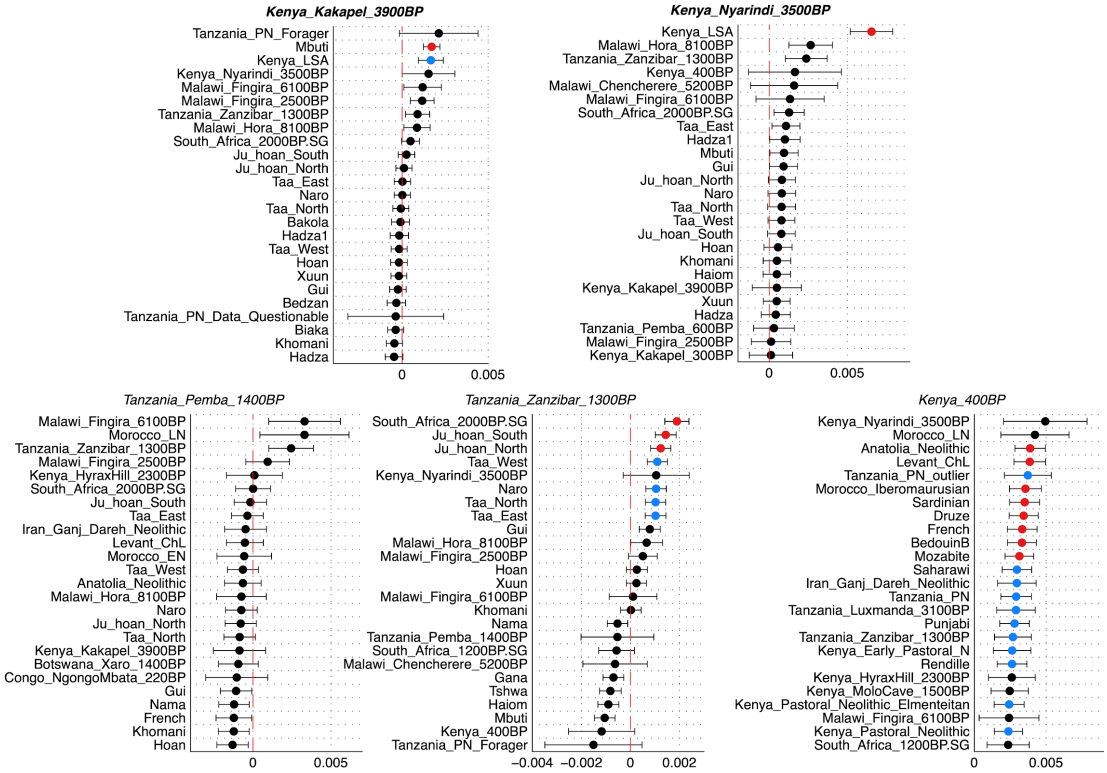


Figure S2. Testing the genetic affinity against Ethiopia_4500BP within the ancient eastern African forager cluster via *f4*-statstistics. For groups/individuals in the ancient eastern African forager cluster, we tested against all populations shown on PCA using symmetric *f4*(ancient east African forager, Ethiopia_4500BP; X, Chimp) and plot 25 groups with top *f4*-value**.** *f4-statistic* tests with Z-score ≥ 3 are shown in red, and tests with 2 ≤ Z-score < 3 are shown in blue.

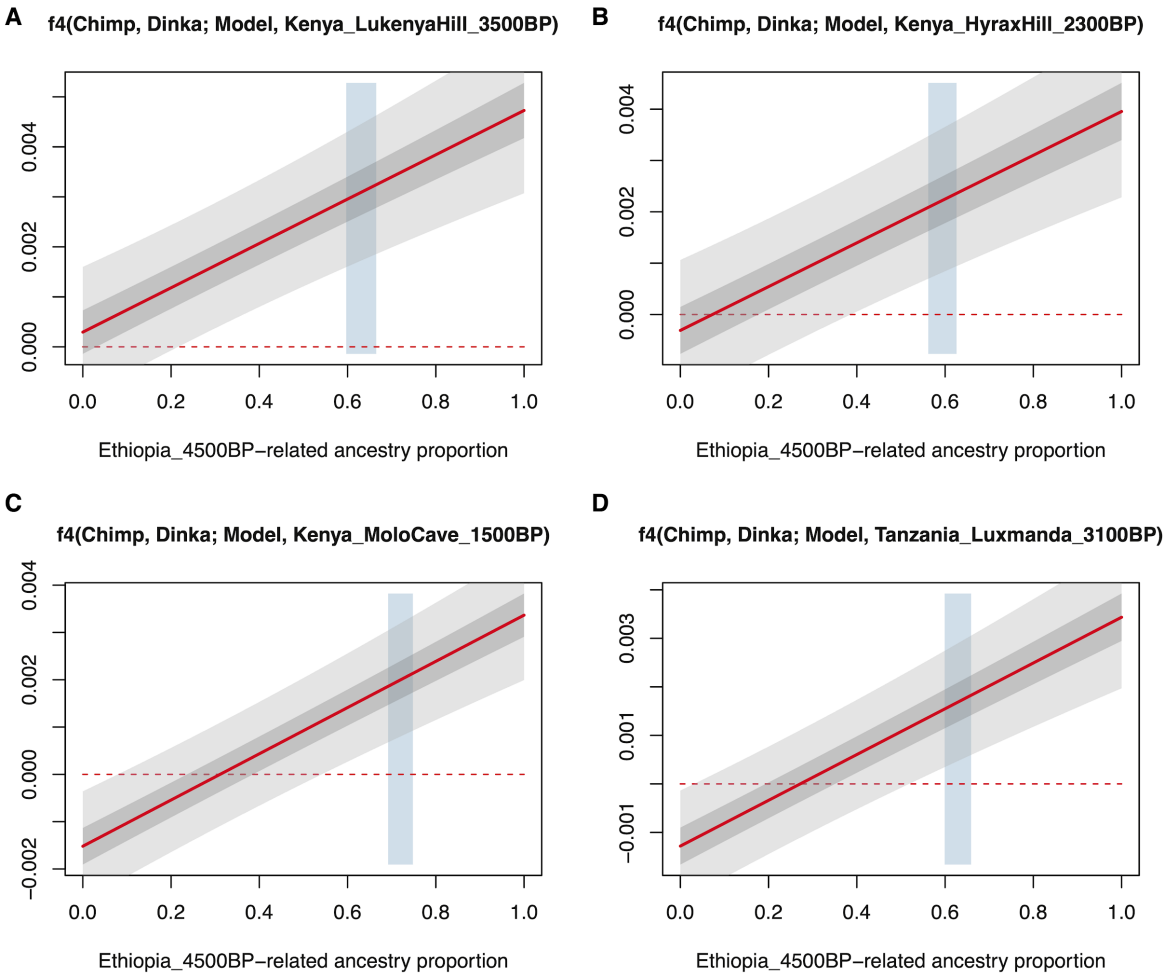


Figure S3. Testing additional genetic affinity to Dinka for ancient eastern African pastoralists based on a two-way admixture model of Ethiopia_4500BP and Levant_ChL . We calculated f4 statistics in form of (Chimpanzee, Dinka; Ethiopia_4500BP+Levant_ChL, ancient Eastern African Pastoralist) for (**A**) Kenya_LukenyaHill_3500BP, (**B**) Kenya_HyraxHill_2300BP, (**C**) Kenya_MoloCave_1500BP, (**D**) Tanzania_Luxmanda_3100BP. Positive F4 values suggests extra affinity to Dinka(above red dashed line), and negative F4 values suggests extra affinity with model. On x-axis, proportions of Ethiopia_4500BP-related ancestry range from 0 to 100% in increments of 0.1%. The blue shade marks the range of ancient Ethiopian ancestry proportion estimated by qpAdm (±1SE). Dark gray and light gray represent ±3 and ±1 SE ranges, respectively. SEs were calculated by 5-centimorgan block jackknife method.

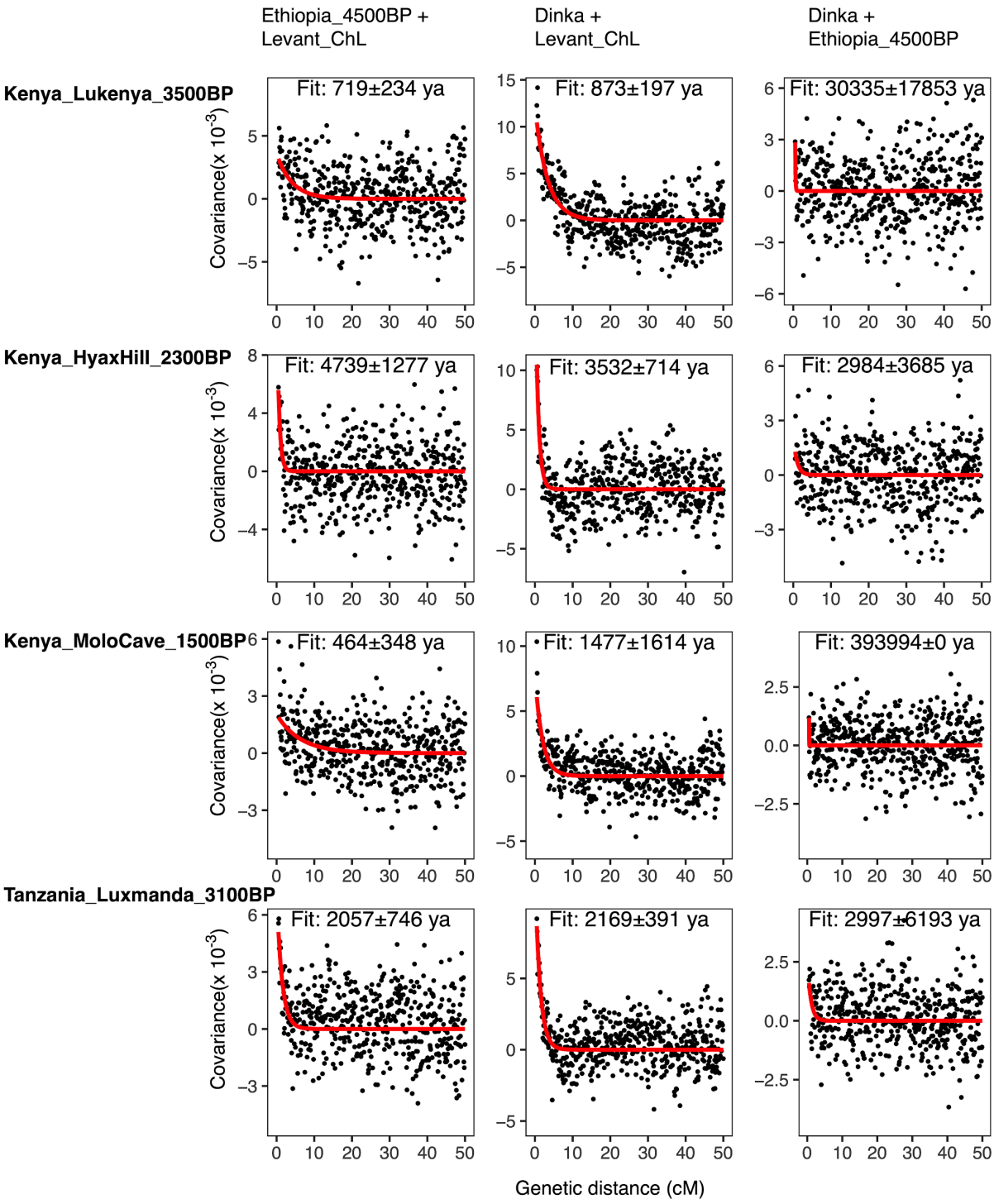


Figure S4. Dating the introduction of ancient Levantine and Dinka ancestry into ancient eastern African pastoralists using DATES. Fitted Linkage Disequilibrium decay curves are shown in red. Bin size of 0.001 Morgans is used for estimation in DATES. Same color and bin size is applied to Fig. S6. We dated three combinations of 2-way admixture from three sources: Ethiopia_4500BP, Levant_ChL, Dinka, shown in different columns, for four target pastoralist groups shown in different rows. DATES-inferred admixture dates between Levant_ChL and two African group show rather reasonable consistent estimates ranging from 2000-7000 years ago. While the admixture date estimates between Dinka and Ethiopia_4500BP are off scale.

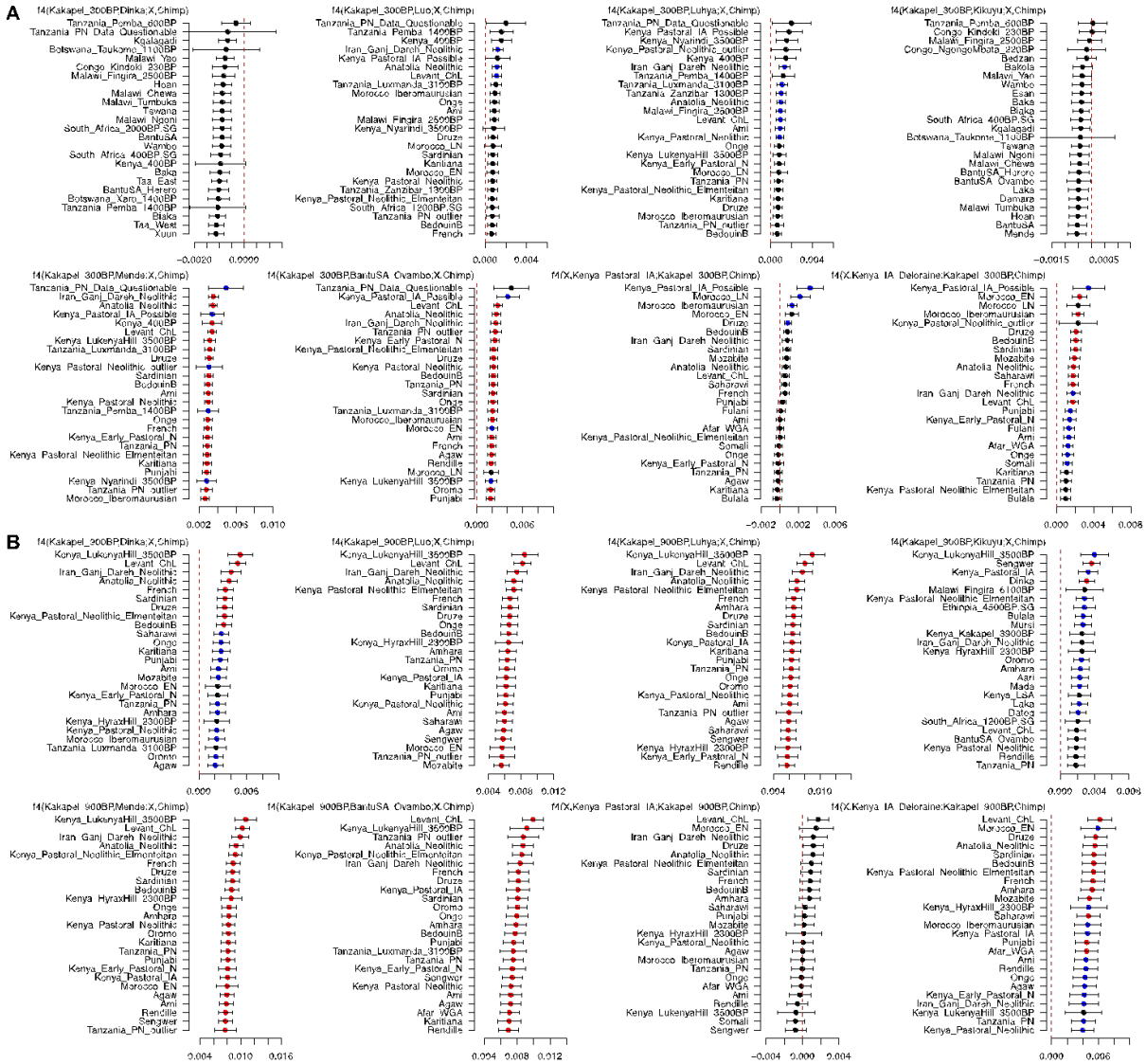


Figure S5. *f-statistics* for ancient individuals from Kakapel site, Kenya. (A) Comparing genetic affinity of Kakapel_300BP to present-day Nilotic-/Bantu-speaking populations and published ancient East African Iron Age genomes. **(B)** Comparing genetic affinity of Kakapel_900BP to present-day Nilotic-/Bantu-speaking populations and published ancient eastern African Iron Age genomes. *f4-statistic* tests with Z-score ≥ 3 are shown in red, and tests with 2 ≤ Z-score < 3 are shown in blue.

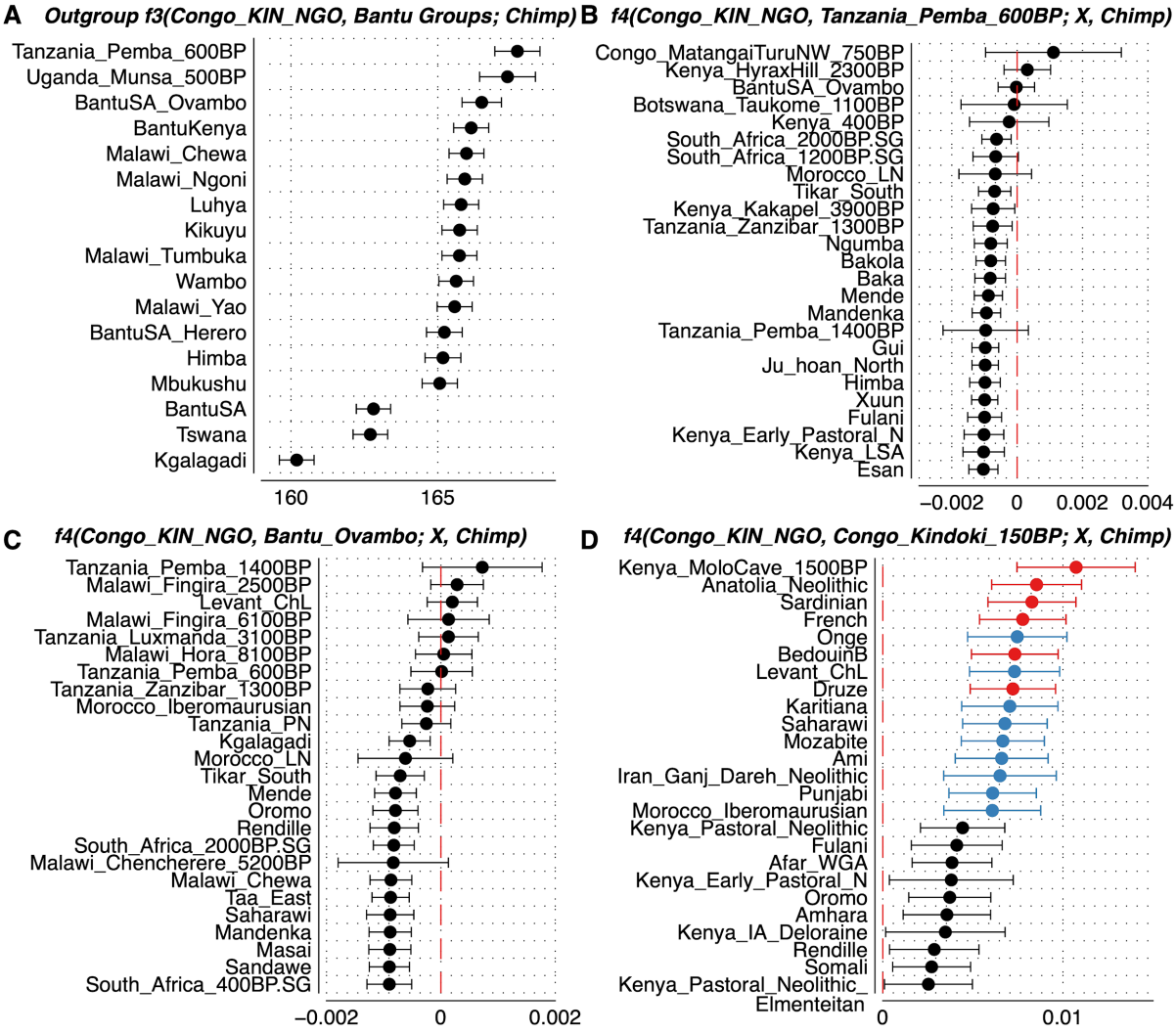
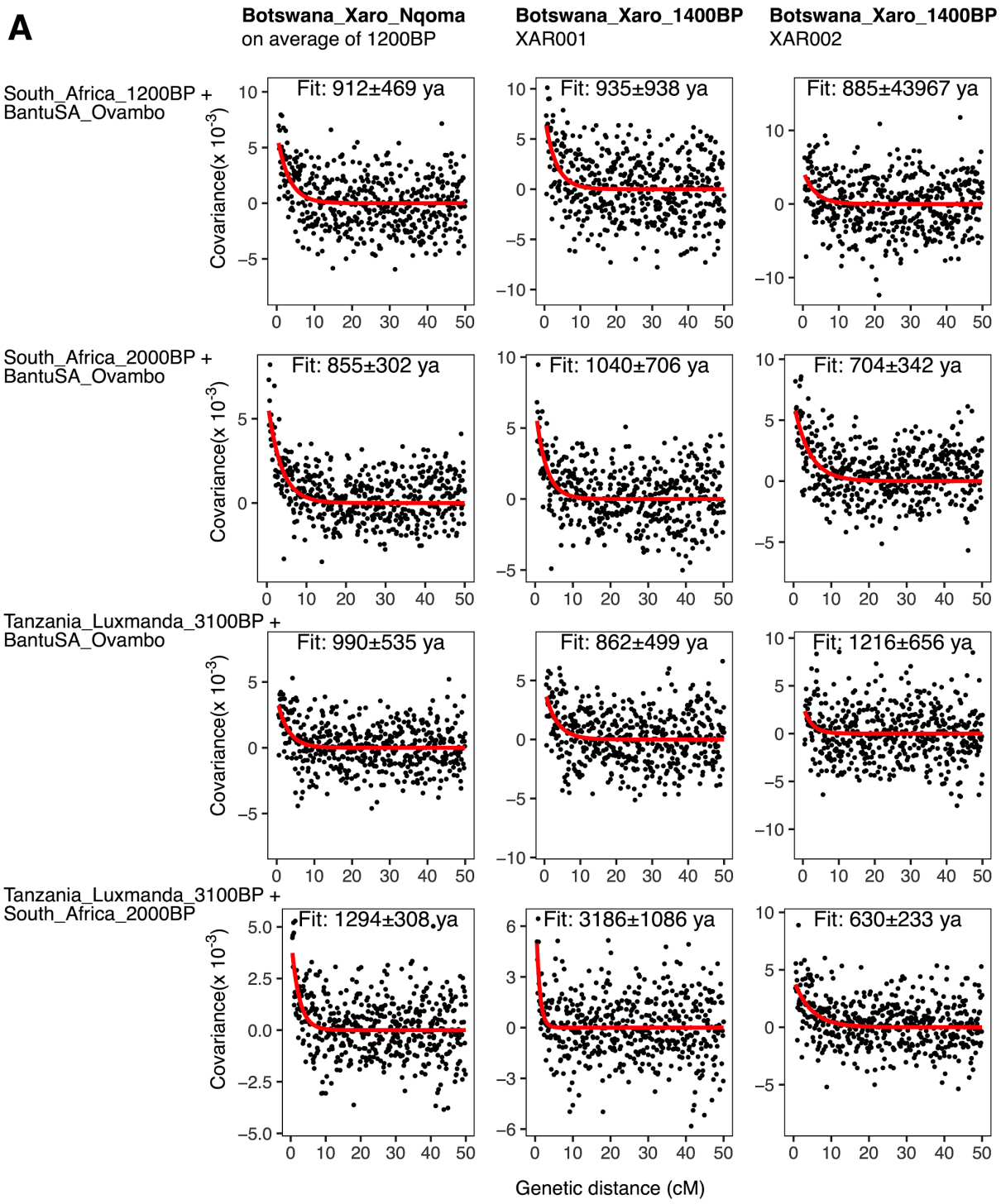


Figure S6. *f-statistics* for ancient individuals from the DR Congo. (A) Comparing the genetic affinity of historical individuals from the DR Congo to ancient Bantu-related groups and present-day Bantu-speaking groups.(B) Testing genetic symmetry between Tanzania_Pemba_600BP and main cluster of historical DR Congo genomes. (C) Testing genetic symmetry between Bantu_Ovambo and the main cluster of historical DR Congo genomes (merged as a single genetic group – "Congo_KIN_NGO"). (D) Cladility *f4* test between Congo_Kindoki_1500BP and the main cluster of historical DR Congo genomes. *f4-statistic* tests with Z-score ≥ 3 are shown in red, and tests with 2 ≤ Z-score < 3 are shown in blue.


A
Botswana_Xaro_Nqoma
on average of 1200BP
Botswana_Xaro_1400BP
XAR001
Botswana_Xaro_1400BP
XAR002
South_Africa_1200BP +
BantuSA_Ovambo
South_Africa_2000BP +
BantuSA_Ovambo
Tanzania_Luxmanda_3100BP +
BantuSA_Ovambo
Tanzania_Luxmanda_3100BP +
South_Africa_2000BP
Fit: 912±469 ya
Fit: 935±938 ya
Fit: 885±43967 ya
Fit: 855±302 ya
Fit: 1040±706 ya
Fit: 704±342 ya
Fit: 990±535 ya
Fit: 862±499 ya
Fit: 1216±656 ya
Fit: 1294±308 ya
Fit: 3186±1086 ya
Fit: 630±233 ya
Covariance(x 10-3)
Genetic distance (cM)

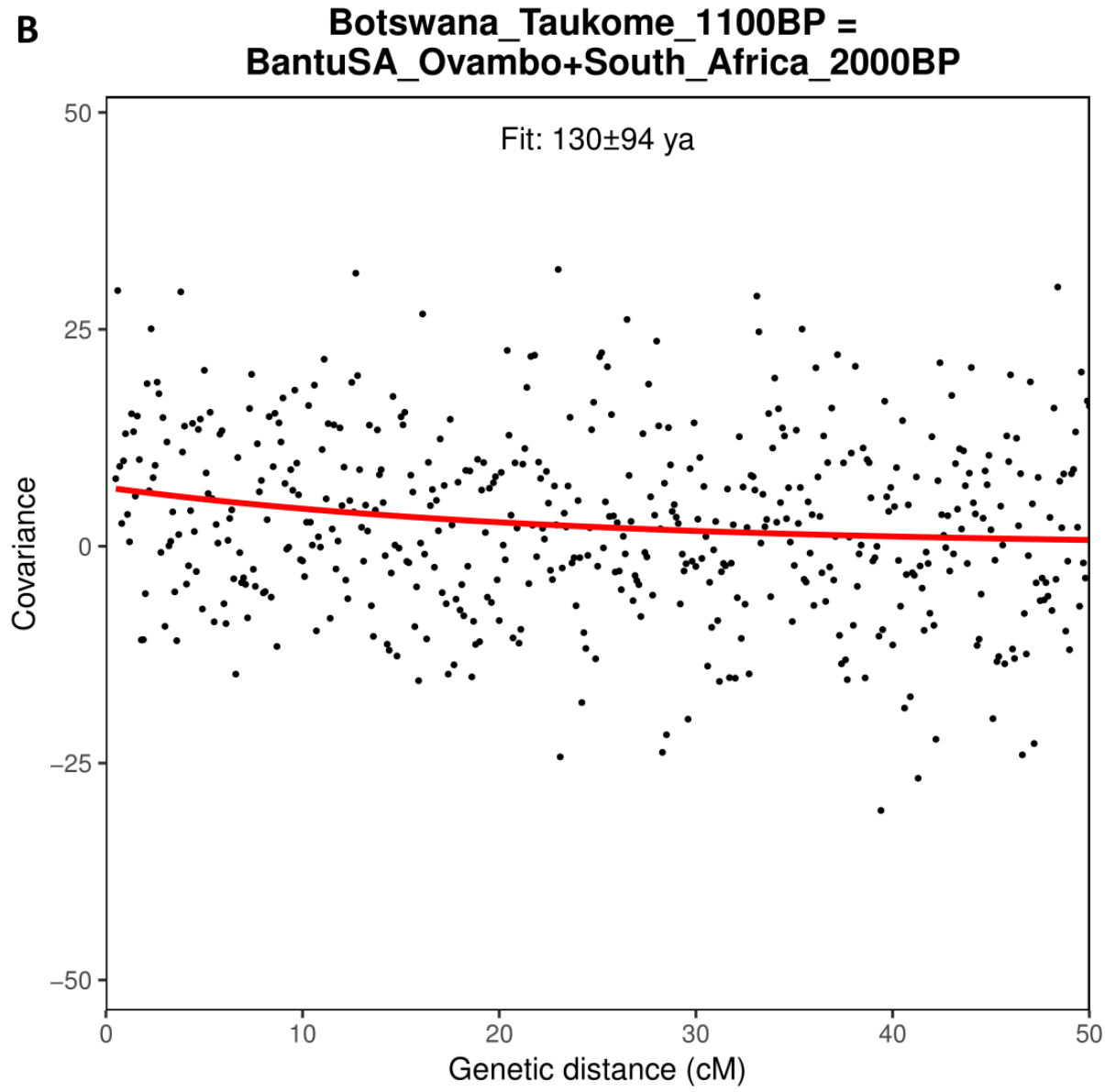


Figure S7. Dating the admixture between Bantu-related ancestry and southern African forager ancestry in ancient Botswana using DATES. In (A), for individuals from Xaro, Nqoma site, we dated four combinations of 2-way admixture from four sources: South_Africa_1200BP, South_Africa_2000BP, BantuSA_Ovambo and Tanzania_Luxmanda_3100BP, shown in different rows, in three ancient Botswana groups (one on grouped-level and the other two on individual-level) shown in different columns. The introduction of Bantu-related ancestry into southern African can be dated to on average of 2,000±400 years ago. The introduction of eastern African pastoralists can be dated into on average of 2500±300 years ago, slightly earlier than the time estimates of incoming Bantu-related ancestry in southern Africa. In (B), for individual Botswana_Taukome_1100BP from Taukome site, we dated the admixture between South_Africa_2000BP and BantuSA_Ovambo occured at 130±94 years ago before the death of this individual.

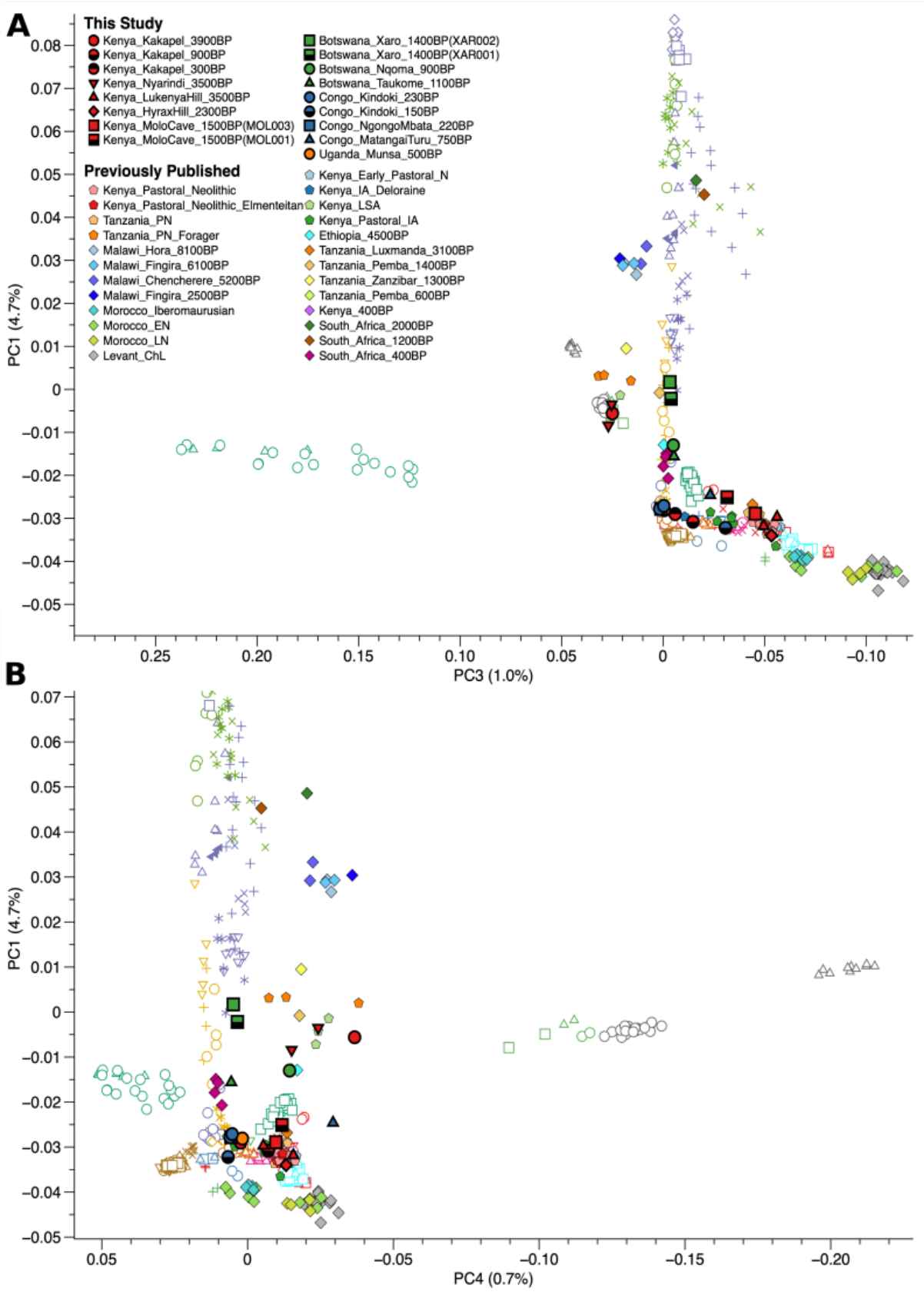


Figure S8. Illustrating genetic variations using more PCs, related to Figure 2 and Fig S2. (A) PC1 versus PC3. (B) PC1 versus PC4. Full legend of modern populations has been shown in Fig S1.

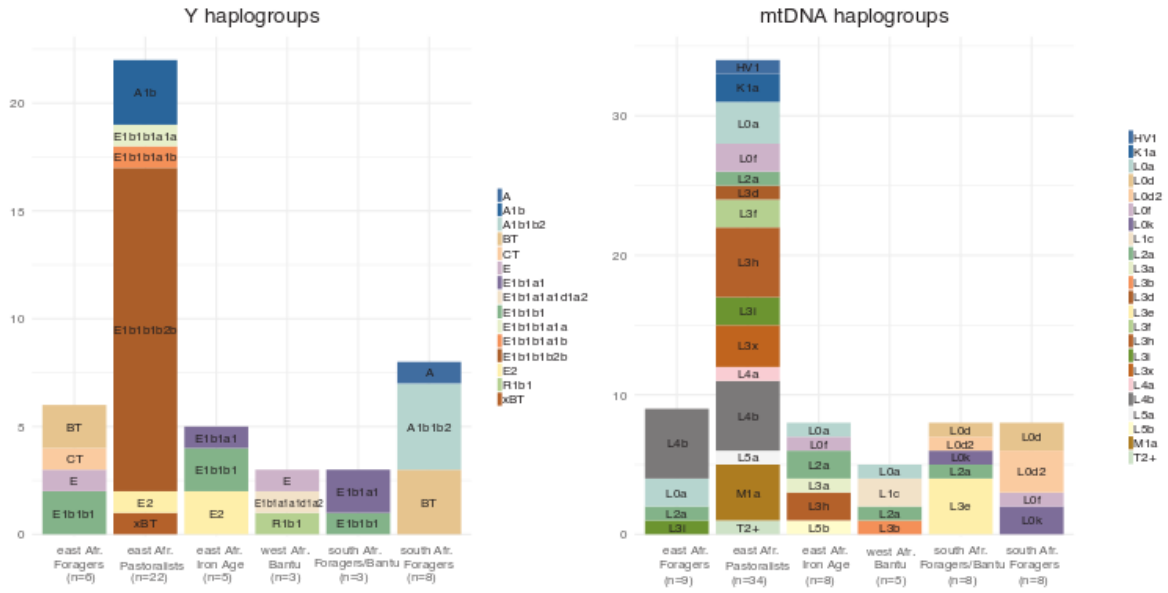


Figure S9. Distribution of mitochondrial and Y chromosome haplogroups in the ancient African genetic clusters. We show (A) the distribution of Y haplogroups in each genetic cluster, (B) distribution of mitochondrial haplogroups in each genetic cluster. The haplogroup information of every genetic cluster is detailed in Table S10.

REFERENCES AND NOTES

1. J. K. Pickrell, N. Patterson, C. Barbieri, F. Berthold, L. Gerlach, T. Güldemann, B. Kure, S. W. Mpoloka, H. Nakagawa, C. Naumann, M. Lipson, P.-R. Loh, J. Lachance, J. Mountain, C. D. Bustamante, B. Berger, S. A. Tishkoff, B. M. Henn, M. Stoneking, D. Reich, B. Pakendorf, The genetic prehistory of southern Africa. *Nat*. *Commun*. **3**, 1143 (2012).
2. C. M. Schlebusch, H. Malmström, T. Günther, P. Sjödin, A. Coutinho, H. Edlund, A. R. Munters, M. Vicente, M. Steyn, H. Soodyall, M. Lombard, M. Jakobsson, Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science*. **358**, 652–655 (2017).
3. P. Skoglund, J. C. Thompson, M. E. Prendergast, A. Mittnik, K. Sirak, M. Hajdinjak, T. Salie, N. Rohland, S. Mallick, A. Peltzer, A. Heinze, I. Olalde, M. Ferry, E. Harney, M. Michel, K. Stewardson, J. I. Cerezo-Román, C. Chiumia, A. Crowther, E. Gomani-Chindebvu, A. O. Gidna, K. M. Grillo, I. T. Helenius, G. Hellenthal, R. Helm, M. Horton, S. López, A. Z. P. Mabulla, J. Parkington, C. Shipton, M. G. Thomas, R. Tibesasa, M. Welling, V. M. Hayes, D. J. Kennett, R. Ramesar, M. Meyer, S. Pääbo, N. Patterson, A. G. Morris, N. Boivin, R. Pinhasi, J. Krause, D. Reich, Reconstructing prehistoric African population structure. *Cell*. **171**, 59–71.e21 (2017).
4. M. E. Prendergast, M. Lipson, E. A. Sawchuk, I. Olalde, C. A. Ogola, N. Rohland, K. A. Sirak, N. Adamski, R. Bernardos, N. Broomandkhoshbacht, K. Callan, B. J. Culleton, L. Eccles, T. K. Harper, A. M. Lawson, M. Mah, J. Oppenheimer, K. Stewardson, F. Zalzala, S. H. Ambrose, G. Ayodo, H. L. Gates Jr, A. O. Gidna, M. Katongo, A. Kwekason, A. Z. P. Mabulla, G. S. Mudenda, E. K. Ndiema, C. Nelson, P. Robertshaw, D. J. Kennett, F. K. Manthi, D. Reich, Ancient DNA reveals a multistep spread of the first herders into sub-Saharan Africa. *Science* **365**, eaaw6275 (2019).
5. R. Pinhasi, D. Fernandes, K. Sirak, M. Novak, S. Connell, S. Alpaslan-Roodenberg, F. Gerritsen, V. Moiseyev, A. Gromov, P. Raczky, A. Anders, M. Pietrusewsky, G. Rollefson, M. Jovanovic, H. Trinhhoang, G. Bar-Oz, M. Oxenham, H. Matsumura, M. Hofreiter, Optimal ancient DNA yields from the inner ear part of the human petrous bone. *PLOS One*. **10**, e0129102 (2015).
6. M. G. Llorente, E. R. Jones, A. Eriksson, V. Siska, K. W. Arthur, J. W. Arthur, M. C. Curtis, J. T. Stock, M. Coltorti, P. Pieruccini, S. Stretton, F. Brock, T. Higham, Y. Park, M.

Hofreiter, D. G. Bradley, J. Bhak, R. Pinhasi, A. Manica, Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science*. **350**, 820–822 (2015).

7. T. Guldemann, A linguist's view: Khoe-Kwadi speakers as the earliest food-producers of southern Africa. *South*. *Afr*. *Humanit*. **20**, 93–132 (2008).
8. J. K. Pickrell, N. Patterson, P.-R. Loh, M. Lipson, B. Berger, M. Stoneking, B. Pakendorf, D. Reich, Ancient west Eurasian ancestry in southern and eastern Africa. *Proc*. *Natl*. *Acad*. *Sci*. *U*.*S*.*A*.**111**, 2632–2637 (2014).
9. M. G. Llorente, E. R. Jones, A. Eriksson, V. Siska, K. W. Arthur, J. W. Arthur, M. C. Curtis, J. T. Stock, M. Coltorti, P. Pieruccini, S. Stretton, F. Brock, T. Higham, Y. Park, M. Hofreiter, D. G. Bradley, J. Bhak, R. Pinhasi, A. Manica, Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. *Science*. **350**, 820–822 (2015).
10. M. van de Loosdrecht, A. Bouzouggar, L. Humphrey, C. Posth, N. Barton, A. Aximu-Petri, B. Nickel, S. Nagel, E. H. Talbi, M. A. El Hajraoui, S. Amzazi, J.-J. Hublin, S. Pääbo, S. Schiffels, M. Meyer, W. Haak, C. Jeong, J. Krause, Pleistocene North African genomes link near Eastern and sub-Saharan African human populations. *Science* **360**, 548–552 (2018).
11. R. Fregel, F. L. Méndez, Y. Bokbot, D. Martín-Socas, M. D. Camalich-Massieu, J. Santana, J. Morales, M. C. Ávila-Arcos, P. A. Underhill, B. Shapiro, G. Wojcik, M. Rasmussen, A. E. R. Soares, J. Kapp, A. Sockell, F. J. Rodríguez-Santos, A. Mikdad, A. Trujillo-Mederos, C. D. Bustamante, Ancient genomes from north Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe. *Proc*. *Natl*. *Acad*. *Sci*. *U*.*S*.*A*.**115**, 6774–6779 (2018).
12. I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, B. Berger, C. Economou, R. Bollongino, Q. Fu, K. I. Bos, S. Nordenfelt, H. Li, C. de Filippo, K. Prüfer, S. Sawyer, C. Posth, W. Haak, F. Hallgren, E. Fornander, N. Rohland, D. Delsate, M. Francken, J.-M. Guinet, J. Wahl, G. Ayodo, H. A. Babiker, G. Bailliet, E. Balanovska, O. Balanovsky, R. Barrantes, G. Bedoya, H. Ben-Ami, J. Bene, F. Berrada, C. M. Bravi, F. Brisighelli, G. B. J. Busby, F. Cali, M. Churnosov, D. E. C. Cole, D. Corach, L. Damba, G. van Driem, S. Dryomov, J.-M. Dugoujon, S. A. Fedorova, I. G. Romero, M. Gubina, M. Hammer, B. M. Henn, T. Hervig, U. Hodoglugil, A. R. Jha, S. Karachanak-Yankova, R. Khusainova, E. Khusnutdinova, R.

Kittles, T. Kivisild, W. Klitz, V. Kučinskas, A. Kushniarevich, L. Laredj, S. Litvinov, T. Loukidis, R. W. Mahley, B. Melegh, E. Metspalu, J. Molina, J. Mountain, K. Näkkäläjärvi, D. Nesheva, T. Nyambo, L. Osipova, J. Parik, F. Platonov, O. Posukh, V. Romano, F. Rothhammer, I. Rudan, R. Ruizbakiev, H. Sahakyan, A. Sajantila, A. Salas, E. B. Starikovskaya, A. Tarekegn, D. Toncheva, S. Turdikulova, I. Uktveryte, O. Utevska, R. Vasquez, M. Villena, M. Voevoda, C. A. Winkler, L. Yepiskoposyan, P. Zalloua, T. Zemunik, A. Cooper, C. Capelli, M. G. Thomas, A. Ruiz-Linares, S. A. Tishkoff, L. Singh, K. Thangaraj, R. Villems, D. Comas, R. Sukernik, M. Metspalu, M. Meyer, E. E. Eichler, J. Burger, M. Slatkin, S. Pääbo, J. Kelso, D. Reich, J. Krause, Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*. **513**, 409–413 (2014).

13. S. Fan, D. E. Kelly, M. H. Beltrame, M. E. B. Hansen, S. Mallick, A. Ranciaro, J. Hirbo, S. Thompson, W. Beggs, T. Nyambo, S. A. Omar, D. W. Meskel, G. Belay, A. Froment, N. Patterson, D. Reich, S. A. Tishkoff, African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biol*.**20**, 82 (2019).
14. S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, D. Reich, The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*. **538**, 201–206 (2016).
15. G. Renaud, V. Slon, A. T. Duggan, J. Kelso, Schmutzi: Estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol*.**16**, 224 (2015).

16. Q. Fu, A. Mittnik, P. L. F. Johnson, K. Bos, M. Lari, R. Bollongino, C. Sun, L. Giemsch, R. Schmitz, J. Burger, A. M. Ronchitelli, F. Martini, R. G. Cremonesi, J. Svoboda, P. Bauer, D. Caramelli, S. Castellano, D. Reich, S. Pääbo, J. Krause, A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr*. *Biol*. **23**, 553–559 (2013).
17. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
18. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. **19**, 1655–1664 (2009).
19. W. Haak, I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, K. Stewardson, Q. Fu, A. Mittnik, E. Bánffy, C. Economou, M. Francken, S. Friederich, R. G. Pena, F. Hallgren, V. Khartanovich, A. Khokhlov, M. Kunst, P. Kuznetsov, H. Meller, O. Mochalov, V. Moiseyev, N. Nicklisch, S. L. Pichler, R. Risch, M. A. Rojo Guerra, C. Roth, A. Szécsényi-Nagy, J. Wahl, M. Meyer, J. Krause, D. Brown, D. Anthony, A. Cooper, K. W. Alt, D. Reich, Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*. **522**, 207–211 (2015).
20. B. M. Henn, L. R. Botigué, S. Gravel, W. Wang, A. Brisbin, J. K. Byrnes, K. Fadhlaoui-Zid, P. A. Zalloua, A. Moreno-Estrada, J. Bertranpetit, C. D. Bustamante, D. Comas, Genomic ancestry of North Africans supports back-to-Africa migrations. *PLOS Genet*. **8**, e1002397 (2012).
21. R. Kuper, S. Kröpelin, Climate-controlled holocene occupation in the sahara: Motor of Africa's evolution. *Science* **313**, 803–807 (2006).
22. A. G. Morris, The myth of the east african `Bushmen'. *South Afr*. *Archaeol*. *Bull*., **58**, 85–90 (2003).
23. T. Güldemann, Greenberg's" case" for Khoisan: The morphological evidence, in *Problems of Linguistic-Historical Reconstruction in Africa*, D. Ibriszimov, Ed. (Köln: Rüdiger Köppe, 2008), pp. 123–153.
24. É. Harney, H. May, D. Shalem, N. Rohland, S. Mallick, I. Lazaridis, R. Sarig, K. Stewardson, S. Nordenfelt, N. Patterson, I. Hershkovitz, D. Reich, Ancient DNA from chalcolithic israel reveals the role of population mixture in cultural transformation. *Nat*. *Commun*. **9**, 3336 (2018).

25. S. H. Ambrose, Chronology of the later stone age and food production in east africa. *J. Archaeol. Sci.***25**, 377–392 (1998).
26. P. J. Lane, The "Moving Frontier" and the transition to food production in Kenya. *Azania* **39**, 243–264 (2004).
27. D. Gifford-Gonzalez, Early pastoralists in east africa: Ecological and social dimensions. *J. Anthropol. Archaeol.* **17**, 166–200 (1998).
28. M. E. Prendergast, K. K. Mutundu, Late holocene archaeological faunas in east Africa: Ethnographic analogues and interpretive challenges. *Documenta Archaeobiologiae*. **7**, 203–232 (2009).
29. J. C. Onyango-Abuje, Crescent island: A preliminary report on excavations at an east african neolithic site. *Azania Archaeol. Res. Africa* **12**, 147–159 (1977).
30. D. P. Gifford, G. L. Isaac, C. M. Nelson, Evidence for predation and pastoralism at prolonged drift: A pastoral neolithic site in kenya. *Azania* **15**, 57–108 (1980).
31. A. B. Smith, Keeping people on the periphery: The ideology of social hierarchies between hunters and herders. *J. Anthropol. Archaeol.* **17**, 201–215 (1998).
32. V. Bajić, C. Barbieri, A. Hübner, T. Güldemann, C. Naumann, L. Gerlach, F. Berthold, H. Nakagawa, S. W. Mpoloka, L. Roewer, J. Purps, M. Stoneking, B. Pakendorf, Genetic structure and sex-biased gene flow in the history of southern african populations. *Am. J. Phys. Anthropol.***167**, 656–671 (2018).
33. B. A. Ogot, *History of the Southern Luo. Volume 1. Migration and Settlement, 1500-1900* (East African Publishing House, 1967).
34. P. Robertshaw, Munsa earthworks: A preliminary report on recent excavations. *Azani Arch. Res. Africa* **32**, 1–20 (1997).
35. S. A. Tishkoff, M. K. Gonder, B. M. Henn, H. Mortensen, A. Knight, C. Gignoux, N. Fernandopulle, G. Lema, T. B. Nyambo, U. Ramakrishnan, F. A. Reed, J. L. Mountain, History of click-speaking populations of africa inferred from mtDNA and Y chromosome genetic variation. *Mol. Biol. Evol.* **24**, 2180–2195 (2007).
36. C. M. Schlebusch, T. Naidoo, H. Soodyall, SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa. *Electrophoresis* **30**, 3657–3664 (2009).
37. B. M. Henn, C. Gignoux, A. A. Lin, P. J. Oefner, P. Shen, R. Scozzari, F. Cruciani, S. A. Tishkoff, J. L. Mountain, P. A. Underhill, Y-chromosomal evidence of a pastoralist

migration through tanzania to southern Africa. *Proc*. *Natl*. *Acad*. *Sci*. *U*.*S*.*A*.**105**, 10693–10698 (2008).

38. N. Isern, J. Fort, Assessing the importance of cultural diffusion in the Bantu spread into southeastern Africa. *PLOS One* **14**, e0215573 (2019).
39. G. Breton, C. M. Schlebusch, M. Lombard, P. Sjödin, H. Soodyall, M. Jakobsson, Lactase persistence alleles reveal partial east african ancestry of southern african Khoe pastoralists. *Curr*. *Biol*. **24**, 852–858 (2014).
40. E. Macholdt, V. Lede, C. Barbieri, S. W. Mpoloka, H. Chen, M. Slatkin, B. Pakendorf, M. Stoneking, Tracing pastoralist migrations to southern Africa with lactase persistence alleles. *Curr*. *Biol*. **24**, 875–879 (2014).
41. N. Sepúlveda, A. Manjurano, S. G. Campino, M. Lemnge, J. Lusingu, R. Olomi, K. A. Rockett, C. Hubbart, A. Jeffreys, K. Rowlands, T. G. Clark, E. M. Riley, C. J. Drakeley; MalariaGEN Consortium, Malaria host candidate genes validated by association with current, recent, and historical measures of transmission intensity. *J Infect Dis*. **216**, 45–54 (2017).
42. K. A. Murphy, A meal on the hoof or wealth in the kraal? Stable isotopes at Kgaswe and Taukome in eastern Botswana. *Int*. *J*. *Osteoarchaeol*. **21**, 591–601 (2011).
43. G. Turner, Early iron age herders in northwestern Botswana: The faunal evidence. *Botsw*. *Notes Rec*. **19**, 7–23 (1987).
44. J. K. Thornton, L. Heywood, Afro-Latino Voices, *Narratives from the Early Modern Ibero-Atlantic World, 1550-1812*, K. J. McKnight, L. J. Garofalo, Eds. (Hackett Publishing, 2009).
45. J. E. Yellen, Barbed bone points: Tradition and continuity in Saharan and sub-Saharan Africa. *African Arch*. *Rev*. **15**, 173–198 (1998).
46. B. Keding, Middle holocene fisher-hunter-gatherers of lake turkana in Kenya and their cultural connections with the north: The pottery. *J*. *African Arch*. **15**, 42–76 (2017).
47. A. Crowther, M. E. Prendergast, D. Q. Fuller, N. Boivin, Subsistence mosaics, forager-farmer interactions, and the transition to food production in eastern Africa. *Quat*. *Int*. **489**, 101–120 (2018).
48. P. Mitchell, Early farming communities of southern and south-central Africa, in *The Oxford Handbook of African Archaeology*, P. Mitchell, P. Lane, Eds. (Oxford Univ. Press, 2013), pp. 657– 670.

49. S. A. Tishkoff, F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A. A. Awomoyi, J.-M. Bodo, O. Doumbo, M. Ibrahim, A. T. Juma, M. J. Kotze, G. Lema, J. H. Moore, H. Mortensen, T. B. Nyambo, S. A. Omar, K. Powell, G. S. Pretorius, M. W. Smith, M. A. Thera, C. Wambebe, J. L. Weber, S. M. Williams, The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044 (2009).
50. P. J. Reimer, E. Bard, A. Bayliss, J. Warren Beck, P. G. Blackwell, C. B. Ramsey, C. E. Buck, H. Cheng, R. Lawrence Edwards, M. Friedrich, P. M. Grootes, T. P. Guilderson, H. Haflidason, I. Hajdas, C. Hatté, T. J. Heaton, D. L. Hoffmann, A. G. Hogg, K. A. Hughen, K. Felix Kaiser, B. Kromer, S. W. Manning, M. Niu, R. W. Reimer, D. A. Richards, E. Marian Scott, J. R. Southon, R. A Staff, C. S. M. Turney, J. van der Plicht, IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon*. **55**, 1869–1887 (2013).
51. C. Bronk Ramsey, T. F. G. Higham, F. Brock, D. Baker, P. Ditchfield, Radiocarbon dates from the Oxford AMS system: *Archaeometry* Datelist 33. *Archaeometry* **51**, 323–349 (2009).
52. J. Dabney, M. Knapp, I. Glocke, M.-T. Gansauge, A. Weihmann, B. Nickel, C. Valdiosera, N. García, S. Pääbo, J.-L. Arsuaga, M. Meyer, Complete mitochondrial genome sequence of a middle pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc*. *Natl*. *Acad*. *Sci*. *U*.*S*.*A*. **110**, 15758–15763 (2013).
53. N. Rohland, E. Harney, S. Mallick, S. Nordenfelt, D. Reich, Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. *Philos*. *Trans*. *R*. *Soc*. *Lond*. *B Biol*. *Sci*. **370**, 20130624 (2015).
54. Q. Fu, M. Hajdinjak, O. T. Moldovan, S. Constantin, S. Mallick, P. Skoglund, N. Patterson, N. Rohland, I. Lazaridis, B. Nickel, B. Viola, K. Prüfer, M. Meyer, J. Kelso, D. Reich, S. Pääbo, An early modern human from Romania with a recent neanderthal ancestor. *Nature* **524**, 216–219 (2015).
55. A. Peltzer, G. Jäger, A. Herbig, A. Seitz, C. Kniep, J. Krause, K. Nieselt, EAGER: Efficient ancient genome reconstruction. *Genome Biol*. **17**, 60 (2016).
56. M. Schubert, S. Lindgreen, L. Orlando, AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC*. *Res*. *Notes* **9**, 88 (2016).
57. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

58. G. Jun, M. K. Wing, G. R. Abecasis, H. M. Kang, An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res*. **25**, 918–925 (2015).
59. H. Weissensteiner, D. Pacher, A. Kloss-Brandstätter, L. Forer, G. Specht, H.-J. Bandelt, F. Kronenberg, A. Salas, S. Schönherr, HaploGrep 2: Mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res*. **44**, W58–W63 (2016).
60. D. Vianello, F. Sevini, G. Castellani, L. Lomartire, M. Capri, C. Franceschi, HAPLOFIND: A new method for high-throughput mtDNA haplogroup assignment. *Hum*. *Mutat*. **34**, 1189–1194 (2013)
61. M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, A. Drummond, Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
62. G. David Poznik, Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. *bioRxiv*, 088716 (2016).
63. J. M. Monroy Kuhn, M. Jakobsson, T. Günther, Estimating genetic kin relationships in prehistoric populations. *PLOS One* **13**, e0195491 (2018).
64. N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, D. Reich, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
65. I. Lazaridis, D. Nadel, G. Rollefson, D. C. Merrett, N. Rohland, S. Mallick, D. Fernandes, M. Novak, B. Gamarra, K. Sirak, S. Connell, K. Stewardson, E. Harney, Q. Fu, G. Gonzalez-Fortes, E. R. Jones, S. A. Roodenberg, G. Lengyel, F. Bocquentin, B. Gasparian, J. M. Monge, M. Gregg, V. Eshed, A.-S. Mizrahi, C. Meiklejohn, F. Gerritsen, L. Bejenaru, M. Blüher, A. Campbell, G. Cavalleri, D. Comas, P. Froguel, E. Gilbert, S. M. Kerr, P. Kovacs, J. Krause, D. McGettigan, M. Merrigan, D. A. Merriwether, S. O'Reilly, M. B. Richards, O. Semino, M. Shamoon-Pour, G. Stefanescu, M. Stumvoll, A. Tönjes, A. Torroni, J. F. Wilson, L. Yengo, N. A. Hovhannisyan, N. Patterson, R. Pinhasi, D. Reich, Genomic insights into the origin of farming in the ancient near east. *Nature*. **536**, 419–424 (2016).
66. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLOS Genet*. **2**, e190 (2006).

67. V. M. Narasimhan, N. Patterson, P. Moorjani, I. Lazaridis, M. Lipson, S. Mallick, N. Rohland, R. Bernardos, A. M. Kim, N. Nakatsuka, I. Olalde, A. Coppa, J. Mallory, V. Moiseyev, J. Monge, L. M. Olivieri, N. Adamski, N. Broomandkhoshbacht, F. Candilio, O. Cheronet, B. J. Culleton, M. Ferry, D. Fernandes, B. Gamarra, D. Gaudio, M. Hajdinjak, É. Harney, T. K. Harper, D. Keating, A. M. Lawson, M. Michel, M. Novak, J. Oppenheimer, N. Rai, K. Sirak, V. Slon, K. Stewardson, Z. Zhang, G. Akhatov, A. N. Bagashev, B. Baitanayev, G. L. Bonora, T. Chikisheva, A. Derevianko, E. Dmitry, K. Douka, N. Dubova, A. Epimakhov, S. Freilich, D. Fuller, A. Goryachev, A. Gromov, B. Hanks, M. Judd, E. Kazizov, A. Khokhlov, E. Kitov, E. Kupriyanova, P. Kuznetsov, D. Luiselli, F. Maksudov, C. Meiklejohn, D. Merrett, R. Micheli, O. Mochalov, Z. Muhammed, S. Mustafokulov, A. Nayak, R. M. Petrovna, D. Pettener, R. Potts, D. Razhev, S. Sarno, K. Sikhymbaeva, S. M. Slepchenko, N. Stepanova, S. Svyatko, S. Vasilyev, M. Vidale, D. Voyakin, A. Yermolayeva, A. Zubova, V. S. Shinde, C. Lalueza-Fox, M. Meyer, D. Anthony, N. Boivin, K. Thangaraj, D. J. Kennett, M. Frachetti, R. Pinhasi, D. Reich, The genomic formation of South and Central Asia. bioRxiv 292581 [**Preprint**] (31 March 2018).
68. M. Feldman, D. M. Master, R. A. Bianco, M. Burri, P. W. Stockhammer, A. Mittnik, A. J. Aja, C. Jeong, J. Krause, Ancient DNA sheds light on the genetic origins of early iron age philistines. *Sci*. *Adv*.. **5**, eaax0061 (2019).
69. C. A. Tryon, I. Crevecoeur, J. T. Faith, R. Ekshtain, J. Nivens, D. Patterson, E. N. Mbua, F. Spoor, Late pleistocene age and archaeological context for the hominin calvaria from GvJm-22 (Lukenya Hill, Kenya). *Proc*. *Natl*. *Acad*. *Sci*. *U*.*S*.*A*. **112**, 2682–2687 (2015).
70. F. Marshall, R. E. B. Reid, S. Goldstein, M. Storozum, A. Wreschnig, L. Hu, P. Kiura, R. Shahack-Gross, S. H. Ambrose, Ancient herders enriched and restructured African grasslands. *Nature* **561**, 387–390 (2018).
71. C. M. Nelson, J. Kimegich, in *Origin and Early Development of Food – Producing Cultures in North-Eastern Africa* (Poznan Archaeological Museum, 1984) pp. 481–487.
72. S. H. Ambrose, M. J. DeNiro, Reconstruction of African human diet using bone collagen carbon and nitrogen isotope ratios. *Nature* **319**, 321–324 (1986).
73. E. A. Sawchuk, thesis, University of Toronto (2017).
74. L. A. Schepartz, thesis, University of Michigan (1987).

75. M. D. Leakey, L. S. B. Leakey, P. M. Game, A. J. H. Goodwin, Report on the excavations at Hyrax Hill, Nakuru, Kenya Colony, 1937–1938. *Trans. R. Soc. S. Afr.* **30**, 271–409 (1943).
76. E. A. Hildebrand, K. M. Grillo, E. A. Sawchuk, S. K. Pfeiffer, L. B. Conyers, S. T. Goldstein, A. C. Hill, A. Janzen, C. E. Klehm, M. Helper, P. Kiura, E. Ndiema, C. Ngugi, J. J. Shea, H. Wang, A monumental cemetery built by eastern Africa's first herders near Lake Turkana, Kenya. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 8942–8947 (2018).
77. H. Field, The University of California African expedition: II, Sudan and Kenya. *Am. Anthropol.* **51**, 72–84 (1949).
78. W. E. Owen, 76. The Early Smithfield culture of Kavirondo (Kenya) and South Africa. *Man.* **41**, 115 (1941).
79. J. L. Buckberry, A. T. Chamberlain, Age estimation from the auricular surface of the ilium: A revised method. *Am. J. Phys. Anthropol.* **119**, 231–239 (2002).
80. E. A. DiGangi, J. D. Bethard, E. H. Kimmerle, L. W. Konigsberg, A new method for estimating age-at-death from the first rib. *Am. J. Phys. Anthropol.* **138**, 164–176 (2009).
81. M. Trotter, R. R. Peterson, Weight of the skeleton during postnatal development. *Am. J. Phys. Anthropol.* **33**, 313–323 (1970).
82. E. C. Lanning, Ancient earthworks in western Uganda. *Uganda J.* **17**, 51–62 (1953).
83. P. Robertshaw, The age and function of ancient earthworks of western Uganda. *Uganda J.* **47**, 20–33 (2001).
84. E. C. Lanning, The munsa earthworks. *Uganda J.* **19**, 177–182 (1955).
85. L. Iles, P. Robertshaw, R. Young, A furnace and associated ironworking remains at Munsa, Uganda. *Azania Arch. Res. Africa* **49**, 45–63 (2014).
86. R. L. Tantala, thesis, University of Wisconsin, Madison (1989).
87. P. Robertshaw, The ancient earthworks of western Uganda: Capital sites of a Cwezi empire? *Uganda J.* **48**, 17–32 (2002).
88. J. Mercader, M. D. Garralda, O. M. Pearson, R. C. Bailey, Eight hundred-year-old human remains from the Ituri tropical forest, democratic republic of congo: The rock shelter site of Matangai Turu northwest. *Am. J. Phys. Anthropol.* **115**, 24–37 (2001).
89. J. Mercader, F. Runge, L. Vrydaghs, H. Doutrelepont, C. E. N. Ewango, J. Juan-Tresseras, Phytoliths from archaeological sites in the tropical forest of Ituri, democratic republic of congo. *Quatern. Res.* **54**, 102–112 (2000).

90. J. Mercader, S. Rovira, P. Gómez-Ramos, Forager-farmer interaction and ancient iron metallurgy in the Ituri rainforest, democratic republic of congo. *Azania Arch*. *Res*. *Africa*. **35**, 107–122 (2000).
91. B. Clist, E. Cranshof, G.-M. de Schryver, D. Herremans, K. Karklins, I. Matonda, C. Polet, A. Sengeløv, F. Steyaert, C. Verhaeghe, K. Bostoen, The elusive archaeology of kongo urbanism: The case of kindoki, Mbanza Nsundi (Lower Congo, DRC). *African Arch*. *Rev*. **32**, 369–412 (2015).
92. B. Clist, E. Cranshof, P. de Maret, M. Kaumba Mazanga, R. Kidebua, I. Matonda, A. Nkanza Lutayi, J. Yogolelo, in *Une Archéologie des Provinces Septentrionales du Royaume Kongo* (Archaeopress, 2018), pp. 135–164.
93. B. Clist, N. Nikis, P. de Maret, in *Une Archéologie des Provinces Septentrionales du Royaume Kongo*, (Archaeopress, 2018), pp. 243–295.
94. J. K. Thornton, in *The Kongo Kingdom: The Origins, Dynamics and Cosmopolitan Culture of an African Polity* (Cambridge Univ. Press, 2018), pp. 17–41.
95. C. Polet, in *Une archéologie des Provinces Septentrionales du Royaume Kongo*, (Archeopress, 2018), pp. 401–438.
96. C. Polet, B.-O. Clist, K. Bostoen, Étude des restes humains de Kindoki (République démocratique du Congo, fin XVIIe–Début XIXe siècle). *Bull*. *Mém*. *Soc*. *Anthropol*. *Paris* **30**, 70–89 (2018).
97. C. Verhaeghe, B.-O. Clist, C. Fontaine, K. Karklins, K. Bostoen, W. De Clercq, Shell and glass beads from the tombs of Kindoki, Mbanza Nsundi, lower congo. *Beads J*. *Soc*. *Bead Res*.. **26**, 23–34 (2014).
98. P. Dubrunfaut, B. Clist, in *Une Archéologie des Provinces Septentrionales du Royaume Kongo*, (Archaeopress, 2018), pp. 359–368.
99. K. Karklins, B. Clist, in *Une Archéologie des Provinces Septentrionales du Royaume Kongo*, (Archaeopress, 2018), pp. 337–348.
100. B. Clist, E. Cranshof, G.-M. de Schryver, D. Herremans, K. Karklins, I. Matonda, F. Steyaert, K. Bostoen, African-European contacts in the Kongo Kingdom (Sixteenth-eighteenth centuries): New archaeological insights from Ngongo Mbata (Lower Congo, DRC). *Int*. *J*. *Hist*. *Archaeol*. **19**, 464–501 (2015).

101. B. Clist, E. Cranshof, M. Kaumba Mazanga, I. Matonda Sakala, A. Nkanza Lutayi, J. Yogolelo, in *Une Archéologie des Provinces Septentrionales du Royaume Kongo* (Archaeopress, 2018), pp. 71–132.

102. M. Bequaert, Fouille d'un cimetière du XVIIe siècle au Congo Belge. *L'Antiquité Classique* **9**, 127–128 (1940).

103. E. Kose, New light on ironworking groups along the middle Kavango in northern Namibia. *South African Arch. Bull.* **64**, 130–147 (2009).

104. M. N. Mosothwane, Dietary stable carbon isotope signatures of the early iron age inhabitants of Ngamiland. *Botsw. Notes Rec.* **43**, 115–129 (2011).

105. E. N. Wilmsen, A. C. Campbell, G. A. Brook, L. H. Robbins, M. Murphy, Mining and moving specular haematite in Botswana, ca. 200–1300 AD, in *The World of Iron* (Archetype, 2013), pp. 33–45.

106. E. N. Wilmsen, Nqoma: An abridged review. *Botsw. Notes Rec.* **43**, 95–114 (2011).

107. J. R. Denbow, E. N. Wilmsen, Iron age pastoralist settlements in Botswana. *S. Afr. J. Sci.* **79**, 405–407 (1983).

108. J. Denbow, thesis, Indiana University (1983).

109. T. N. Huffman, *Handbook to the Iron Age* (Univ. of KwaZulu-Natal Press, 2007).

14.3. Supplementary Materials of paper C

CellPress
OPEN ACCESS

Supplemental Figures

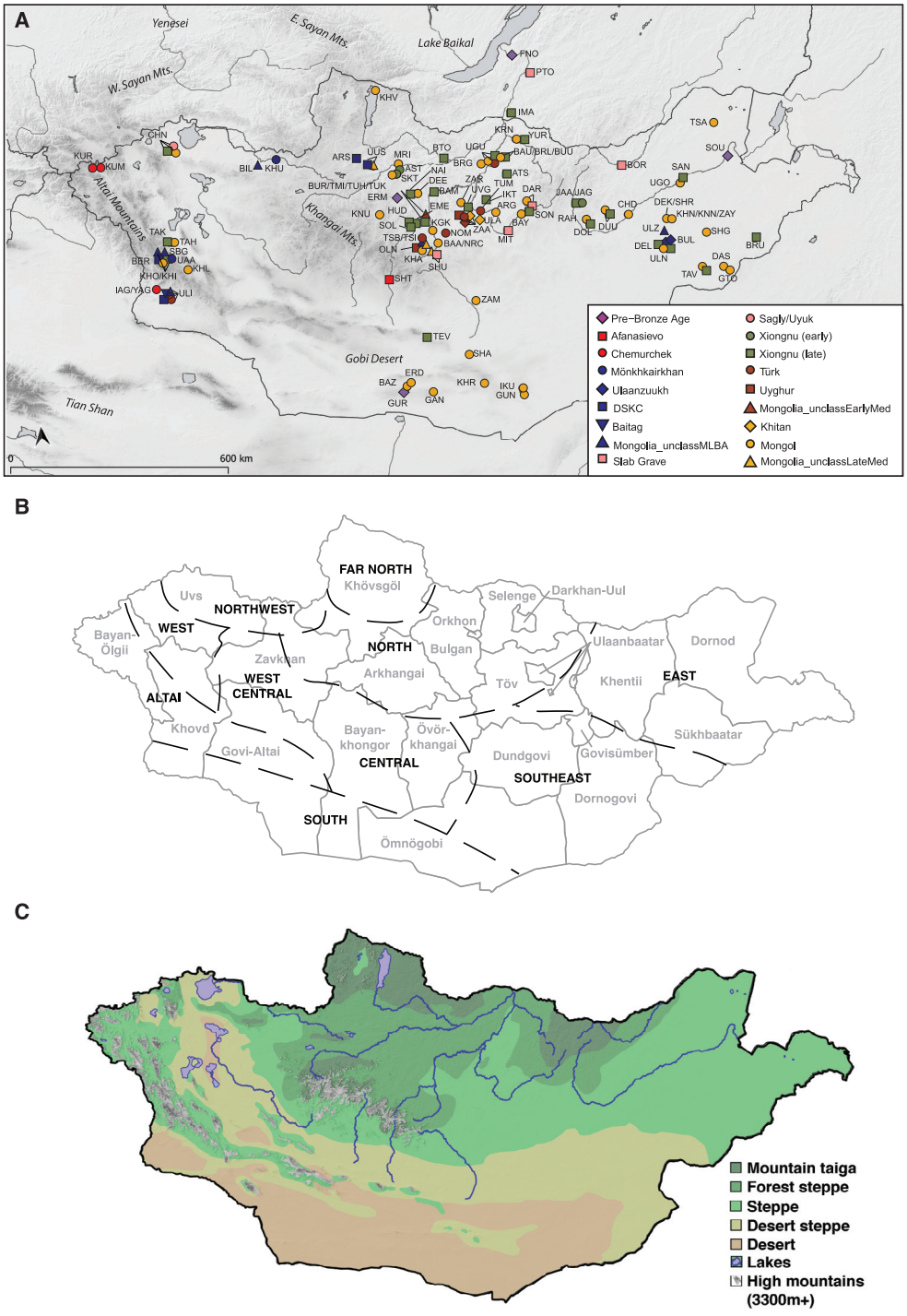


Figure S1. Archaeological Sites and Geographic and Ecological Features in Mongolia, Related to Figure 1

(A) Archaeological sites in Mongolia and neighboring regions analyzed in this study.

(B) Mongolian regions and provinces (aimags). Provinces are indicated by gray lines and text. Regions are indicated by black dashed lines and text following the definitions of (Taylor et al., 2019).

(C) Ecological zones of Mongolia. Map produced using QGIS software (v3.6) with ecological data from (Dorjgotiv, 2004).

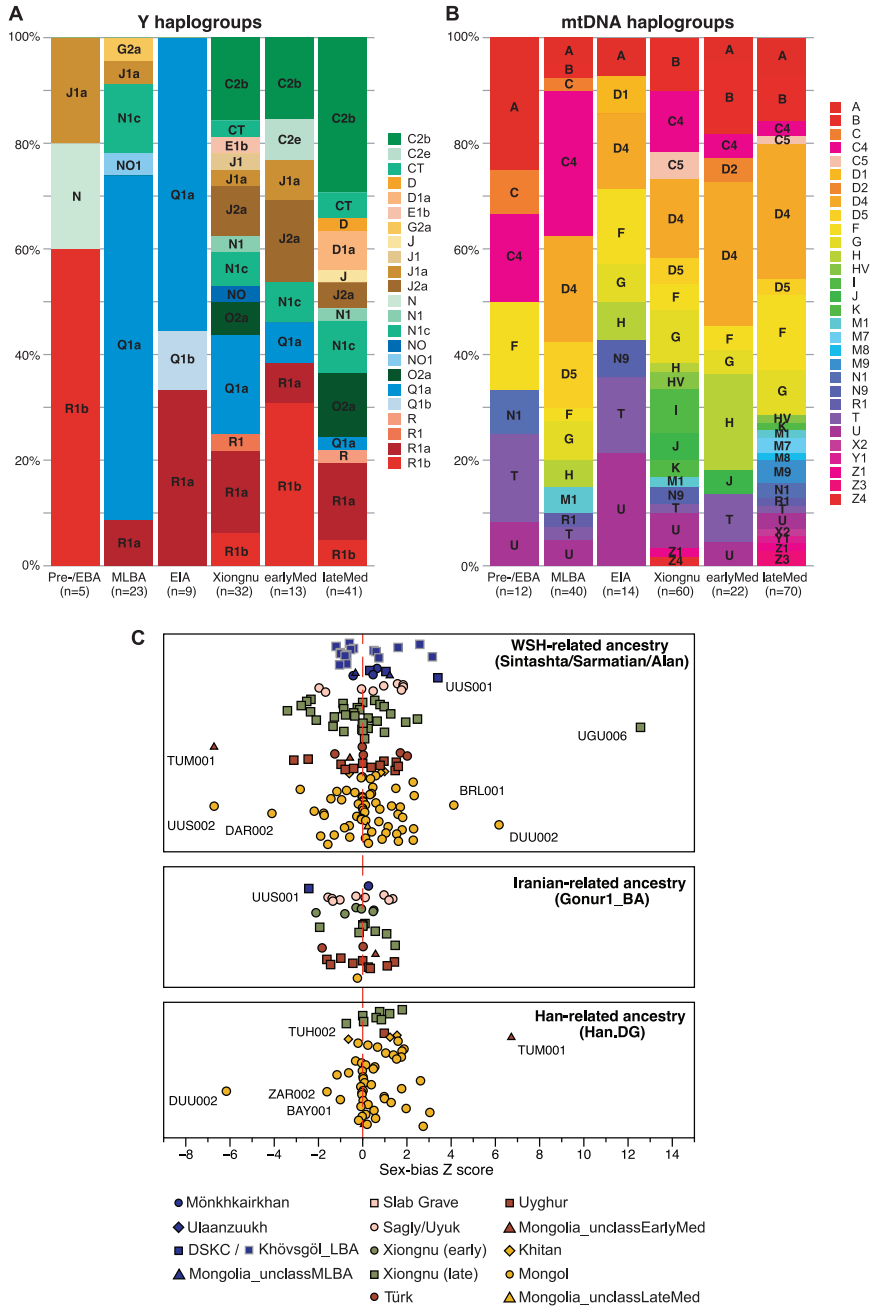
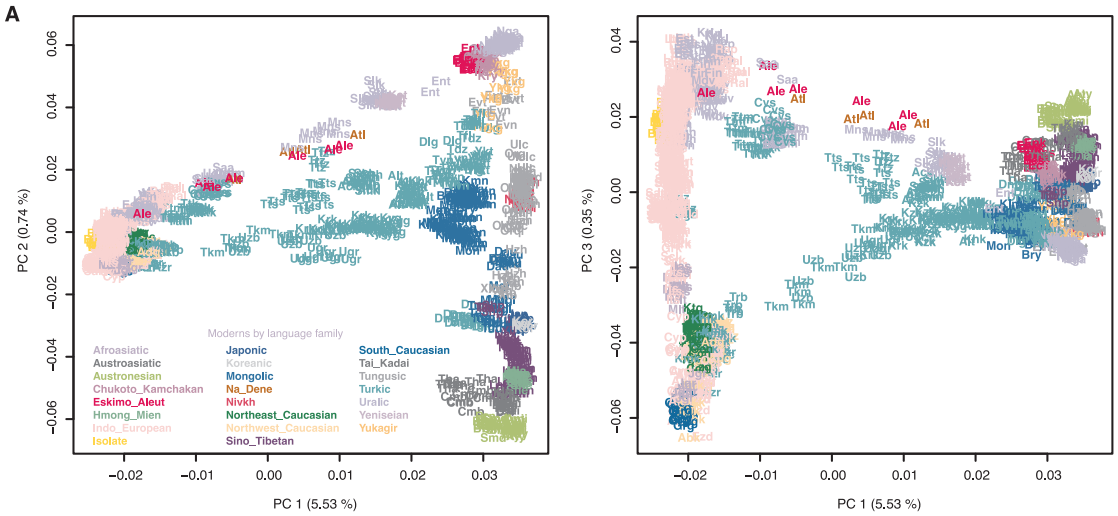


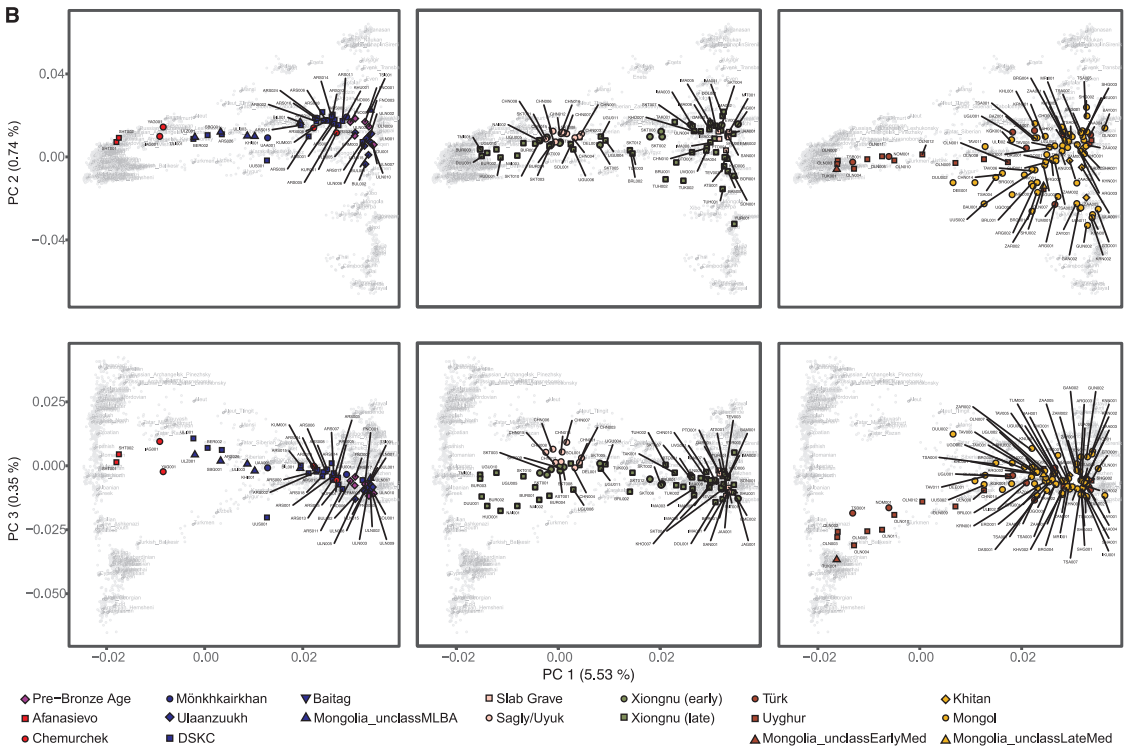
Figure S2. Uniparental Haplogroup Assignments by Group and Sex-Bias *Z* Scores, Related to Figure 5B and STAR Methods

(A and B) Population structure from uniparentally inherited markers. (A) Distribution of Y haplogroups across each period. (B) Distribution of mitochondrial haplogroups across each period.

(C) Sex-bias Z scores by evaluating the differences of WSH-/Iranian-/Han-related ancestry on the autosomes and the X chromosome. We calculated Z-score for each ancient individual who has genetic admixture with any of the three ancestries. Positive scores suggest more WSH-/Iranian-/Han-related ancestry on the autosomes, i.e., male-driven admixture.



Abz Abazin
Abk Abkhasian
Adg Adygei
Alb Albanian
Ale Aleut
Atl Aleut_Tlingit
Alt Altaian
Ack Altaian_Chelkans
Ami Ami
Arm Armenian
Ahm Armenian_Hemsheni
Aty Atayal
Avr Avar
Azr Azeri
Blk Balkar
Bsc Bashkir
Bsq Basque
Blr Belarusian
Bes Besermyan
Brn Borneo
Blg Bulgarian
Bry Buryat
Cmb Cambodian
Cvs Chuvash
Ccs Circassian
Crt Croatian
Cyp Cypriot
Cze Czech
Dai Dai
Drg Darginian
Dau Daur
Dlg Dolgan
Dng Dungan
Ent Enets
Eng English
Ecs Eskimo_CS
Enk Eskimo_NK
Est Estonian
Evn Even
Evf Evenk_FarEast
Evt Evenk_Transbaikal
Ezd Ezid
Fin Finnish
Fra French
Ggz Gagauz
Grg Georgian
Han Han
Hzh Hezhen
Hun Hungarian
Ice Icelandic
Igs Ingushian
Iri Irish
Iru Irish_Ulster
Itn Italian_North
Its Italian_South
Ite Itelmen
Jap Japanese
Jas Jew_Ashkenazi
Jgr Jew_Georgian
Kbd Kabardinian
Ktg Kaitag
Klm Kalmyk
Krc Karachai
Krk Karakalpak
Krl Karelian
Kzk Kazakh
Ket Ket
Khs Khakass
Khk Khakass_Kachins
Kor Korean
Kry Koryak
Kbc Kubachinian
Kmk Kumyk
Kvd Kuvd
Kyg Kyrgyz
Lah Lahu
Lak Lak
Lzg Lezgin
Lth Lithuanian
Mlt Maltese
Mns Mansi
Mia Miao
Mld Moldavian
Mon Mongol
Mgl Mongola
Mdv Mordovian
Nan Nanai
Nax Naxi
Ngd Negidal
Nga Nganasan
Nvh Nivh
Ng1 Nogai
Orq Oroqen
Ost Ossetian
Pol Polish
Rom Romanian
Rus Russian
Rak Russian_Krasnoborsky
Ral Russian_Leshukonsky
Rap Russian_Pinezhsky
Saa Saami
Sar Sardinian
Sct Scottish
Slk Selkup
Smd Semende
She She
Shp Sherpa
Sll Shetlandic
Skh Shor_Khakassia
Smn Shor_Mountain
Scl Sicilian
Srb Sorb
Spa Spanish
Spn Spanish_North
Tbs Tabasaran
Ttm Tatar_Mishar
Tts Tatar_Siberian
Ttz Tatar_Zabolotniye
Tha Thai
Tib Tibetan
Tdz Todzin
Tfl Tofalar
Tuu Tu
Tbl Tubalar
Tuj Tujia
Tra Turkish
Trb Turkish_Balikesir
Tkm Turkmen
Tvn Tuvinian
Udm Udmurt
Ukr Ukrainian
Ulc Ulchi
Ugr Uygur
Uzb Uzbek
Vep Veps
Xib Xibo
Ykt Yakut
Yii Yi



(legend on next page)

Figure S3. PCA of Present-Day Eurasian Populations and Genetic Structure of Mongolia through Time, Related to Figure 2

(A) PCA of present-day Eurasian populations used as the background for Figure 2 and Figure S3B. Here we show the population labels for the 2,077 Eurasian individuals used for calculating PCs and plotted as gray dots in Figure 2. Each three-letter code in the plot represents a single individual. Population IDs matching to the three-letter codes are listed at the bottom.

(B) Genetic structure of Mongolia through time. Principal component analysis (PCA) of ancient individuals (n = 214) from three major periods projected onto contemporary Eurasians (gray symbols). Projection and axis variance corresponds to Figure 2. Population labels are positioned over the mean coordinate across individuals belonging to each population.

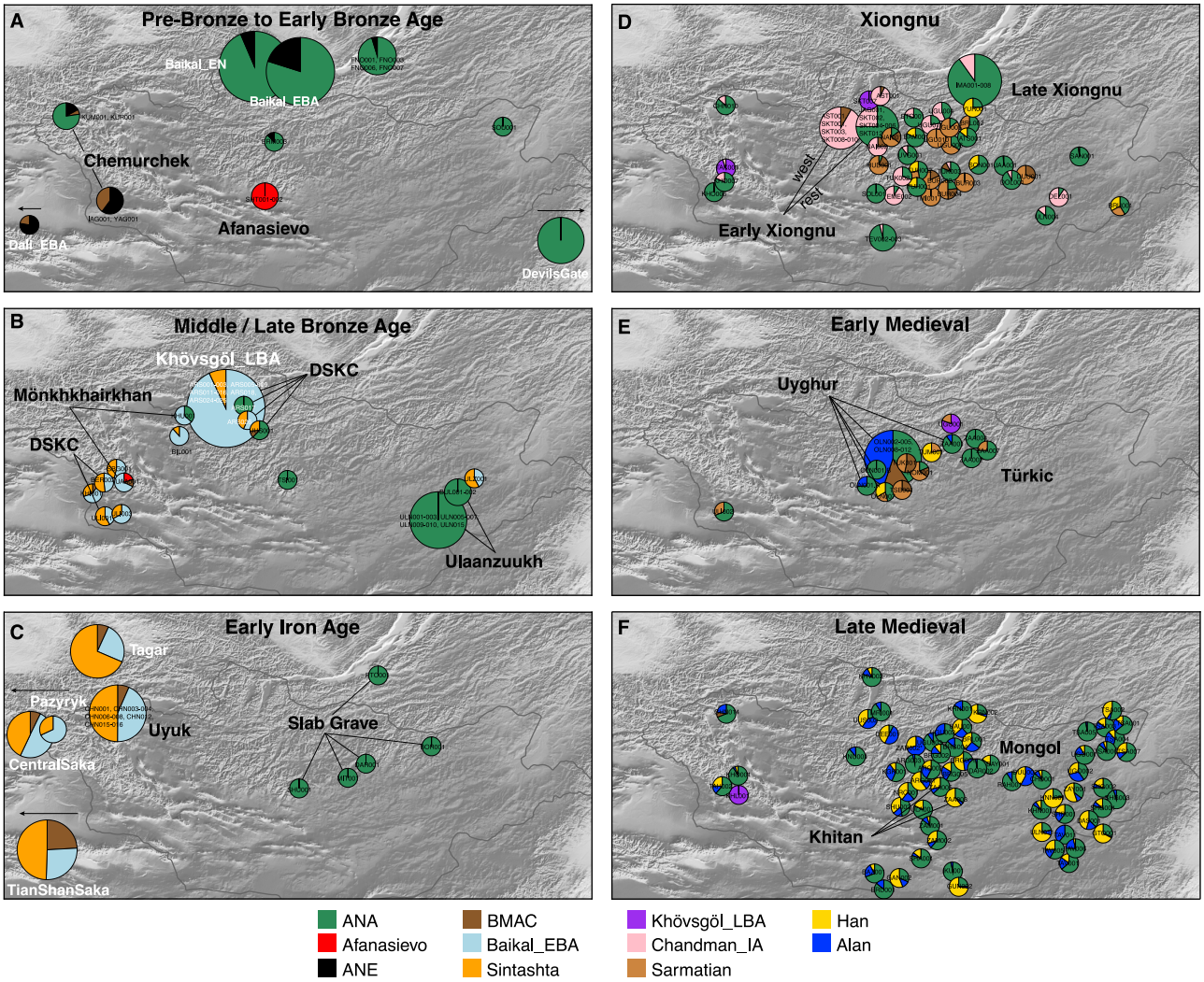


Figure S4. Genetic Changes in the Eastern Steppe across Time Characterized by qpAdm with All Individuals Indicated, Related to Figures 3 and 4

(A) Pre-Bronze through Early Bronze Age;
(B) Middle/Late Bronze Age;
(C) Early Iron Age;
(D) Xiongnu period;
(E) Early Medieval;
(F) Late Medieval.

Modeled ancestry proportions are indicated by sample size-scaled pie charts, with ancestry source populations shown below. Cultural groups are indicated by bold text. For panels (D–F), individuals are Late Xiongnu, Türkic, and Mongol, respectively, unless otherwise noted. Previously published reference populations are noted with white text; all others are from this study. Populations beyond the map borders are indicated by arrows. Burial locations have been jittered to improve visibility of overlapping individuals. Zoom in to see individual labels. Here we report results from admixture models that include all ancestry components required to explain historic late Medieval individuals as a group for unbiased cross comparison between individuals. Individual results with simpler admixture models can be found in Table S5J. See modeling details in Section 7.

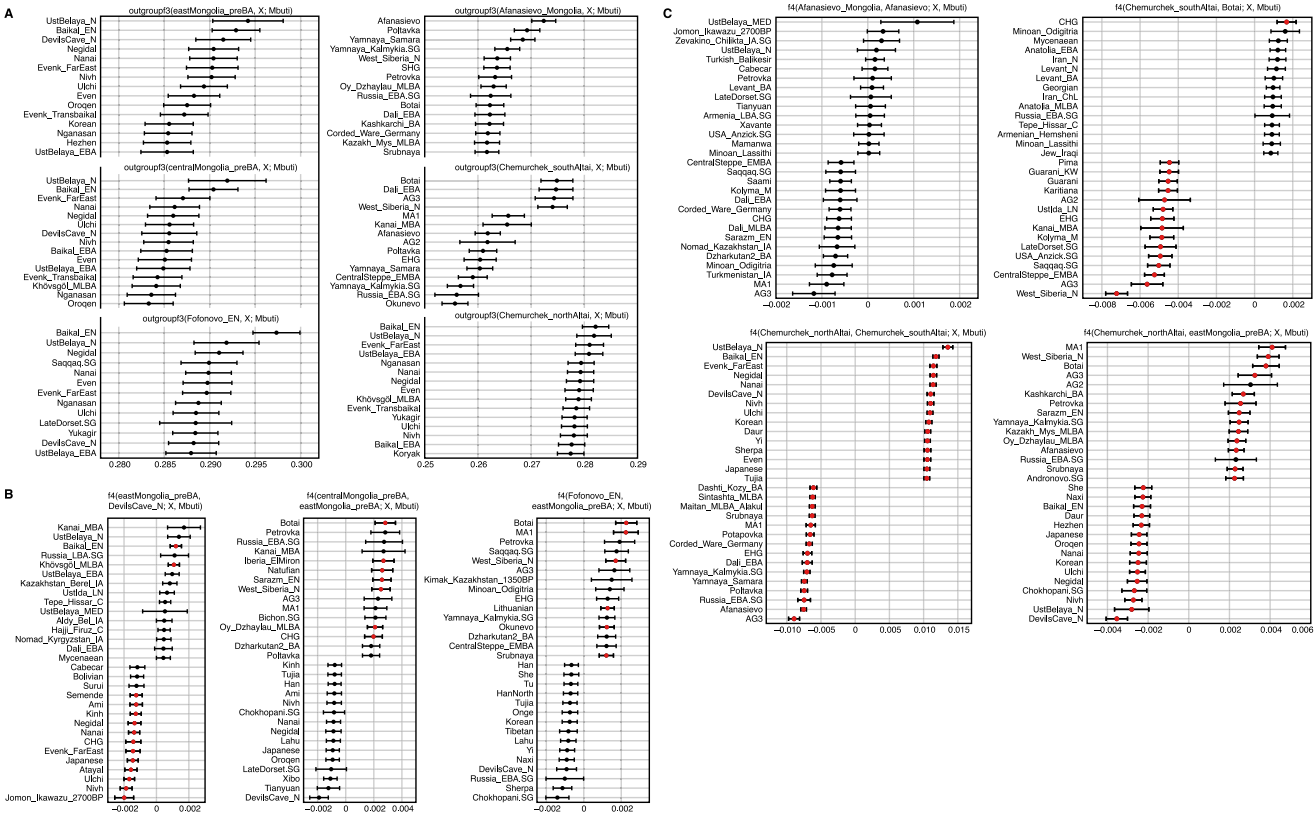
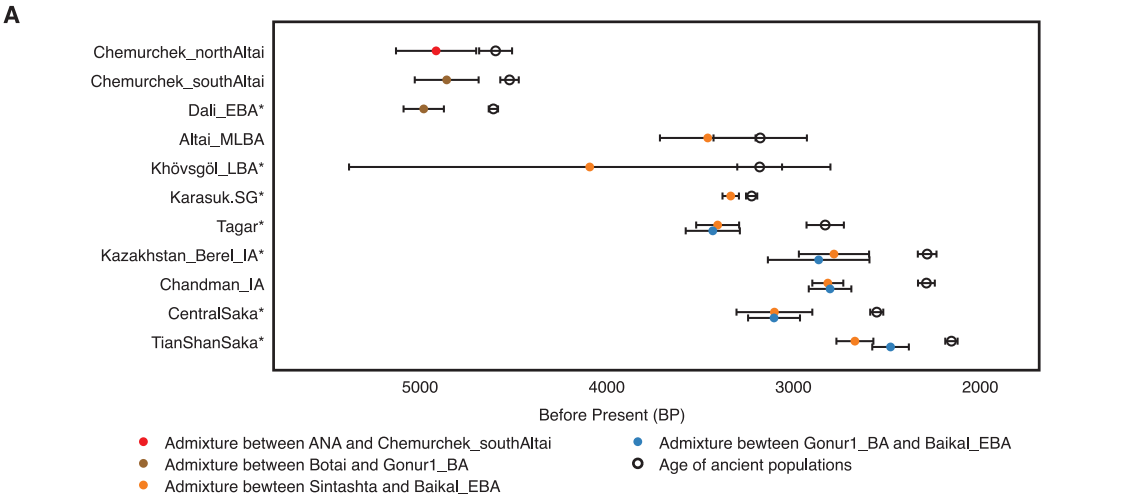


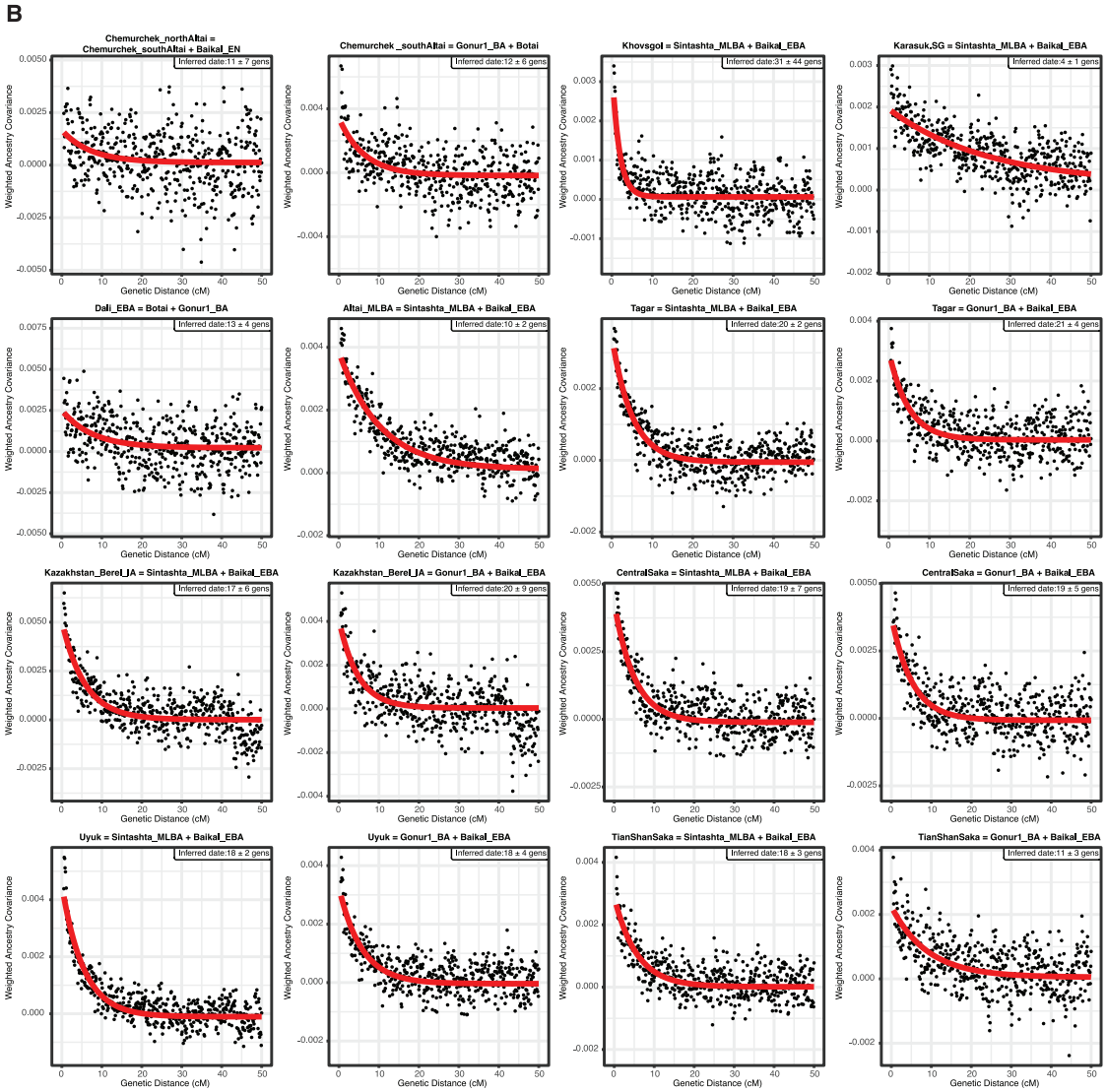
Figure S5. Outgroup *f3*-Statistics and Cladality Testing using f4-Statistics, Related to Figures 3 and 4

(A) Outgroup *f3*-statistics for the pre-Bronze Age to Early Bronze Age groups in the Eastern Steppe. We show top 15 outgroup *f3*-statistics of the form *f3*(Target, world-wide; Mbuti) out of 345 ancient and present-day populations for the six target groups: eastMongolia_preBA, centralMongolia_preBA, Fofonovo_EN, Afanasievo_Mongolia, Chemurchek_southAltai and Chemurchek_northAltai. Horizontal bars represent ± 1 standard error (SE) calculated by 5 cM block jackknifing.

(B) Testing cladality of the four ANA populations using *f4*-statistics. We show top and bottom 15 symmetric *f4*-statistics of the form *f4*(ANA1, ANA2; world-wide, Mbuti) out of 345 ancient and present-day populations for the four ANA-related target groups: eastMongolia_preBA, centralMongolia_preBA, Fofonovo_EN, DevilsCave_N. Horizontal bars represent ± 1 standard error (SE) calculated by 5 cM block jackknifing. *f4*-statistics with Z-score > 3 are highlighted in red.

(C) Testing cladality of Afanasievo and Chemurchek using *f4*-statistics. We show top and bottom 15 symmetric *f4*-statistics for the three target groups Afanasievo_Mongolia, Chemurchek_southAltai and Chemurchek_northAltai, in the form *f4*(Afanasievo_Mongolia, Afanasievo; world-wide, Mbuti), *f4*(Chemurchek_southAltai, Botai; world-wide, Mbuti), *f4*(Chemurchek _northAltai, Chemurchek_southAltai; world-wide, Mbuti), and *f4*(Chemurchek _northAltai, eastMongolia_preBA; world-wide, Mbuti) out of 345 ancient and present-day populations. Horizontal bars represent ± 1 standard error (SE) calculated by 5 cM block jackknifing. *f4*-statistics with Z-score > 3 are highlighted in red.





(legend on next page)

Figure S6. Dating Admixture in Prehistoric Individuals, Related to STAR Methods

(A) Dating admixture in prehistoric individuals. We estimated admixture dates using the DATES program and converted it by adding the age of each ancient population (mean value of the center of the 95% confidence interval of calibrated ^{14}C dates) and assuming 29 years per generation. Horizontal bars associated with the admixture dates (colored circles) are estimated by the square root of summing the variance of DATES estimate using leave-one-chromosome-out jackknifing method and the variance of the ^{14}C date estimate, assuming that the two quantities are independent. Published groups are marked with an asterisk (*). For the Chemurchek_northAltai, we used Baikal_EN as the representative of ANA ancestry for dating the admixture event, given the larger sample size of Baikal_EN.

(B) Ancestry covariance in prehistoric individuals. We show the weighted ancestry covariance (y axis) calculated from DATES which is expected to decay exponentially along genetic distance (x axis) with a decay rate indicating the time since admixture, and fitted exponential curves (shown in red line). We start the fit at genetic distance at 0.45 centiMorgans, and estimate standard error by a weighted block jackknife removing one chromosome in each run.

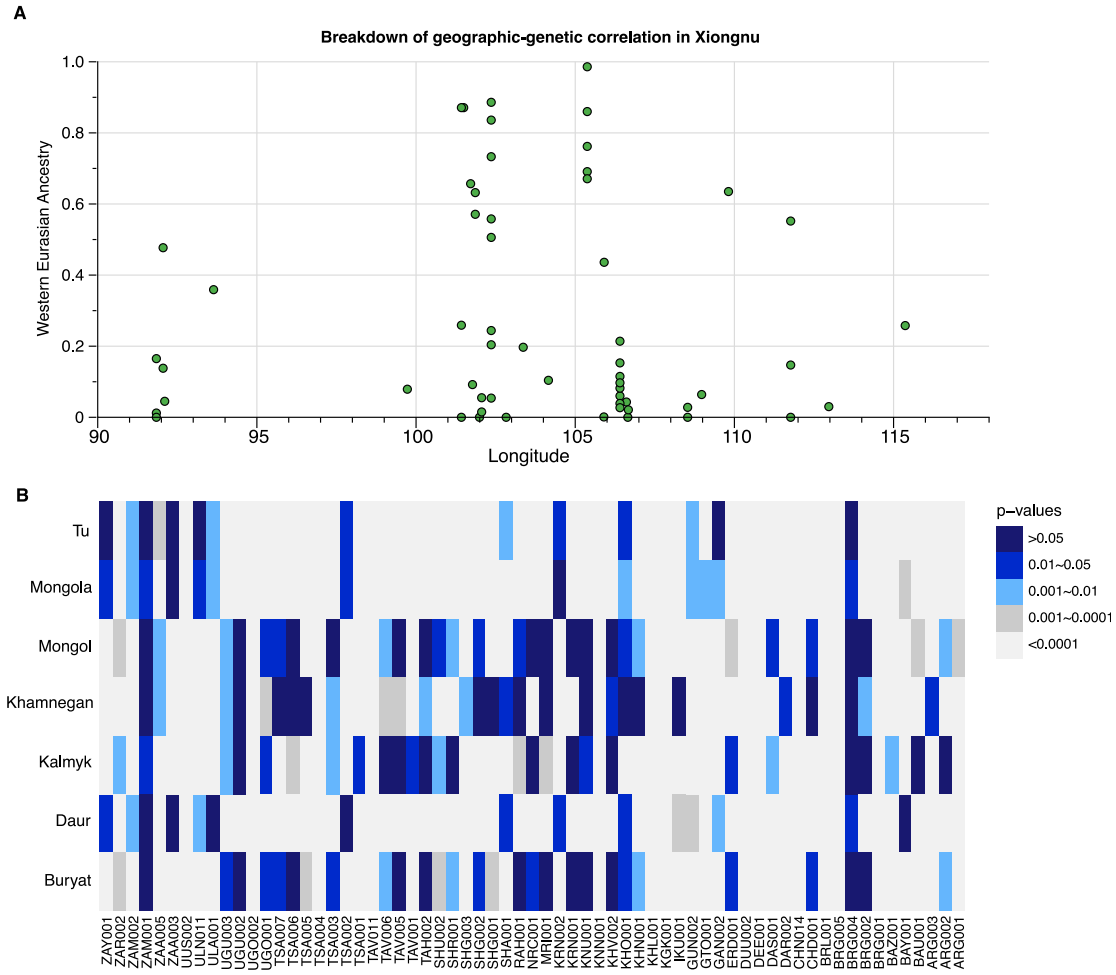


Figure S7. Breakdown of Geography and Genetics among Xiongnu and Comparison of Mongol Period and Present-Day Populations, Related to Figure 3 and STAR Methods

(A) Breakdown of the geographic-genetic correlation in Xiongnu. We show the proportions of West Eurasian ancestry on all individuals/groups from Xiongnu era (y axis) versus the longitude of archaeological site they come from (x axis). The raw numbers of individual estimates can be found in Table S5G for models using Sarmatian as the western Eurasian source. Unlike MLBA/EIA individuals (Figure 3), Xiongnu individuals from more western sites do not have higher proportion of western Eurasian ancestry than those from eastern sites.

(B) Comparing genetic homogeneity between ancient Mongol individuals and seven present-day Mongolic-speaking populations using qpWave. We report the *p-value* for every individual-based qpWave {ancient Mongol individual; Mongolic group} using seven modern Mongolic-speaking populations: Buryat, Daur, Kalmyk, Khamnegan, Mongol, Mongola, and Tu in the Human Origins dataset. When the *p-value* from qpWave is > 0.05, it suggests that the ancient individual on the y axis is genetically indistinguishable from the modern Mongolic-speaking population shown on the x axis. Smaller *p-value*s indicate that the ancient individual is significantly different from the modern group.